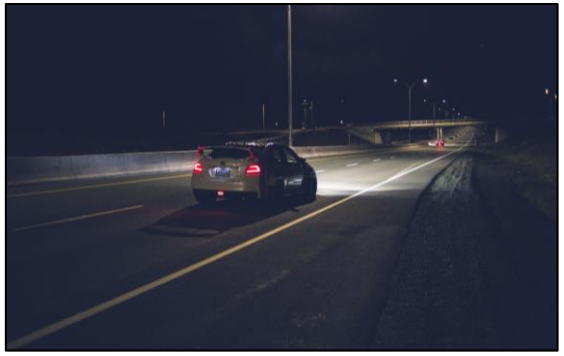
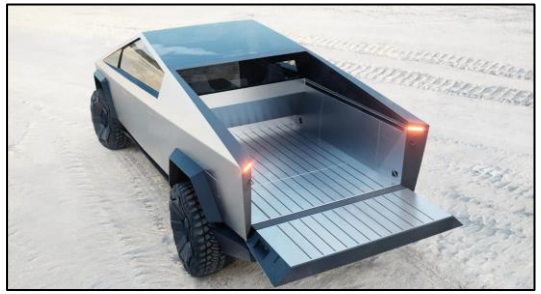
Hierarchical Dataset Curation & Adaptative Learning, Towards Robust 2D/3D Computer Vision

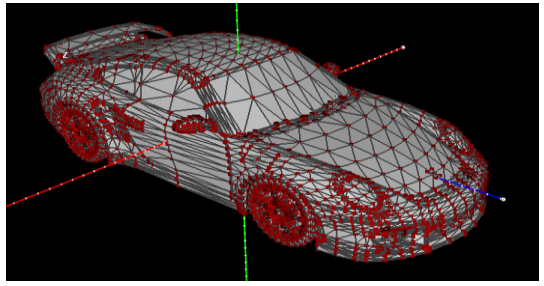
Brandon Leung

Tesla Interview Presentation
Friday, June 10, 2022

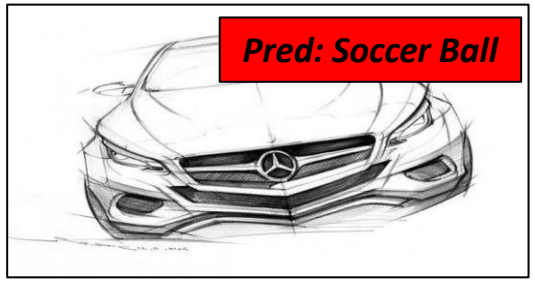
What is Invariance & Robustness in Vision?

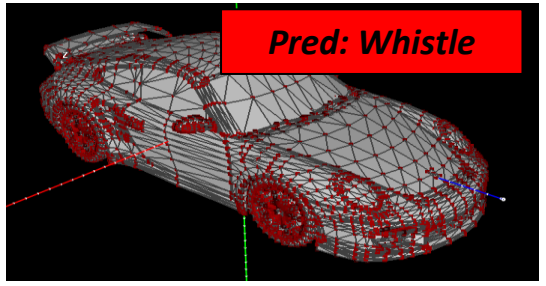






Invariance of EfficientNet Trained on ImageNet

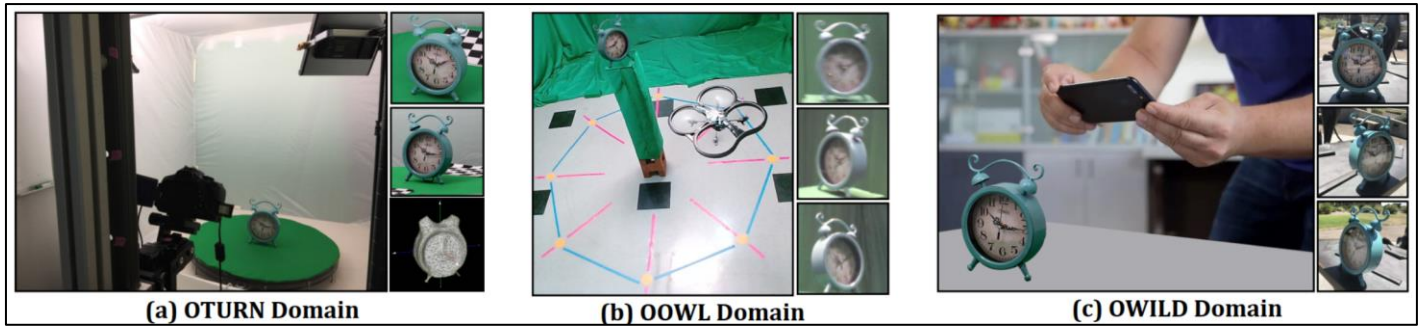


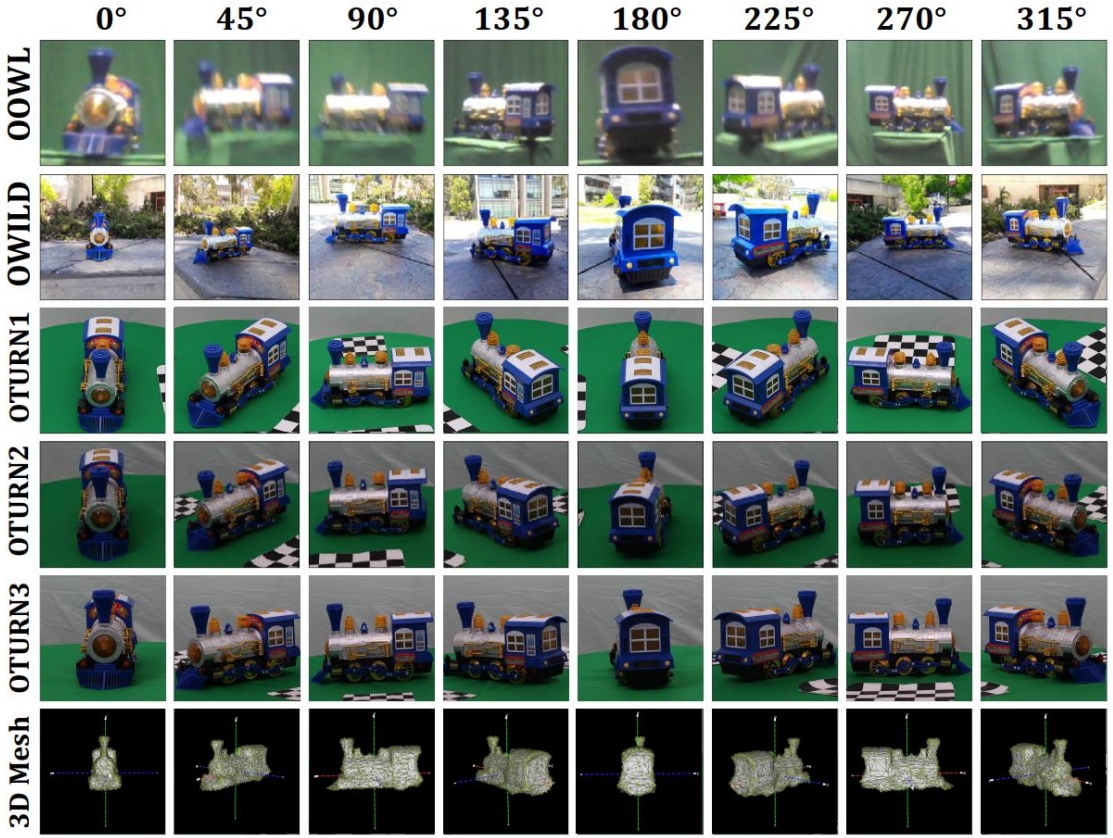


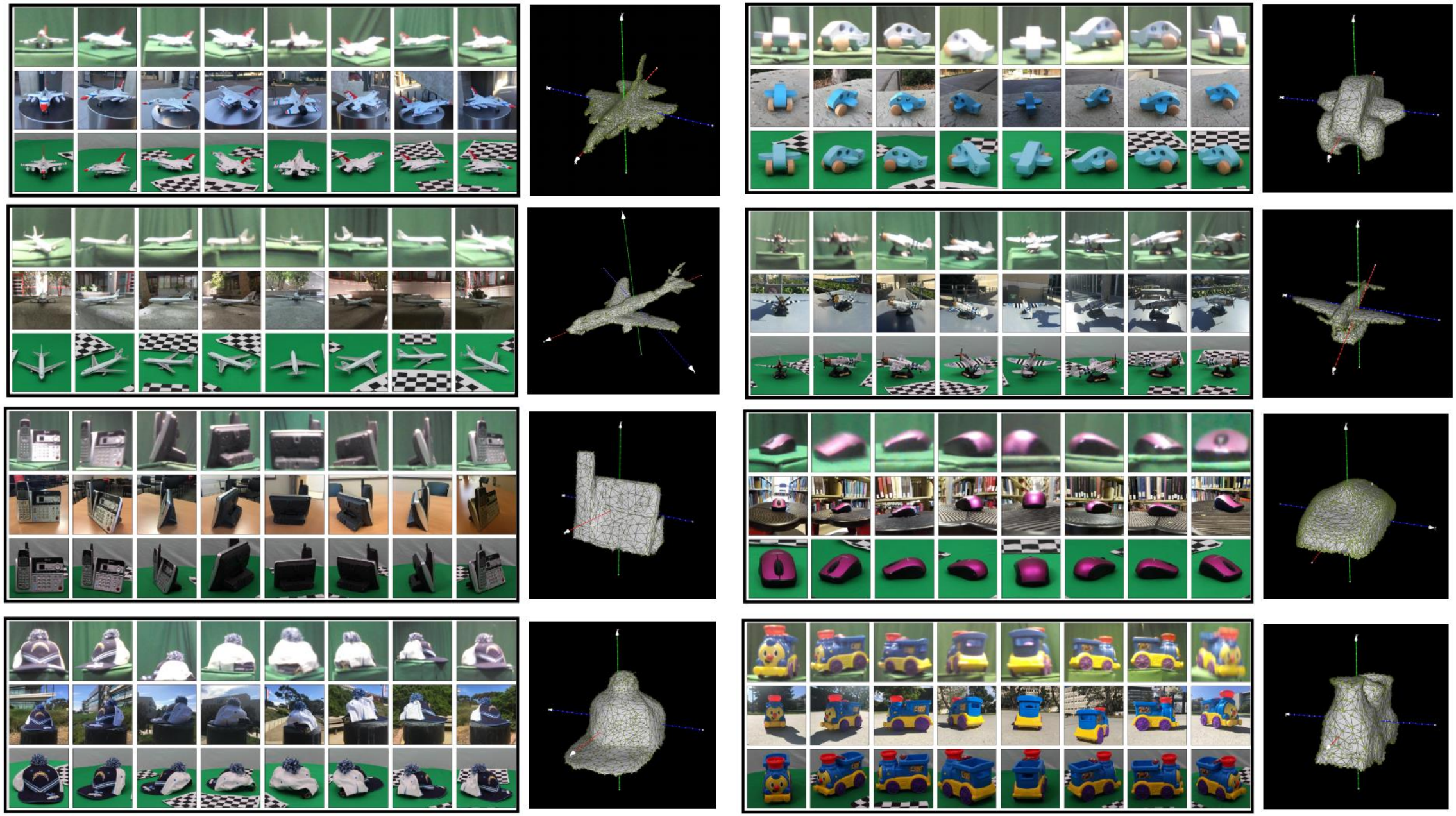
- **Difficult for CNNs to generalize** across domains without extensive labeled data
- ImageNet's test set alone cannot test for true invariance, and leaves "**blind spots**" from data bias undiscovered

3D Object Domain Dataset Suite (3D-ODDS)

- Carefully designed, hierarchical real-world dataset with **~200K images** and **331 3D meshes**
- 3 disentangled variation factors: **Domain, Viewpoint, Class**
- ~25 classes, ~20 obj instances per class
- Ideal as a **benchmark** or **testing framework** for invariance



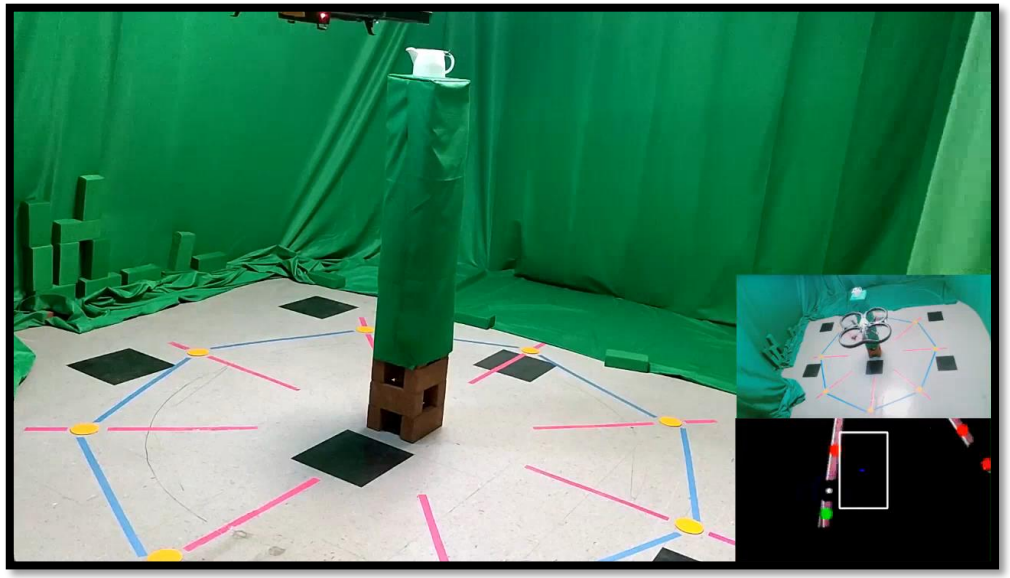




Data Collection Behind the Scenes

Led 13 Contributors in 3 Subgroups:

- Drone Flight Dev.
- Dataset Collection
- ML Research & Experiments



Drone flight space

Dataset objects

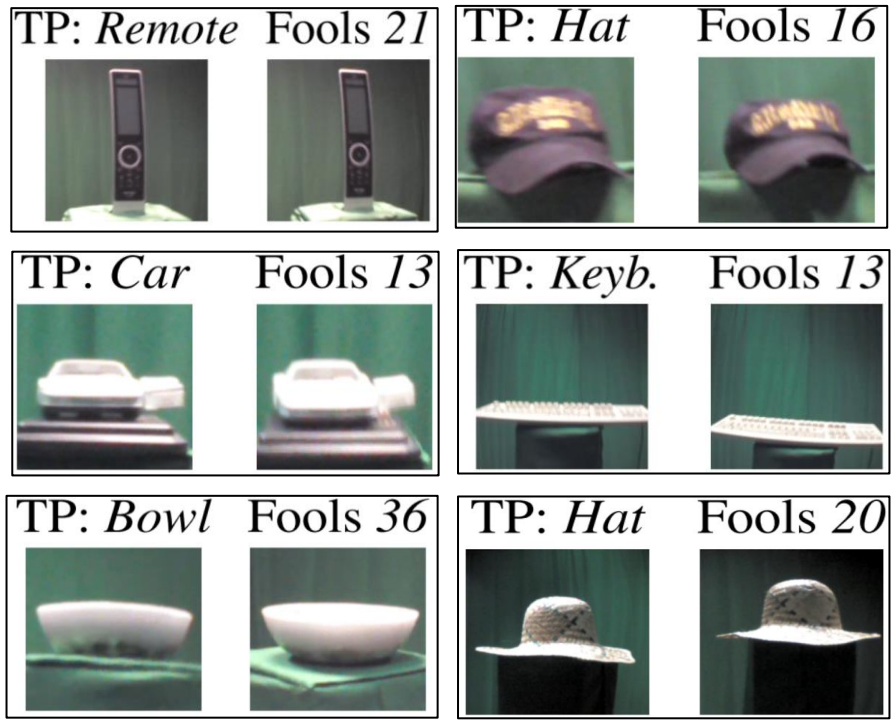
More dataset objects

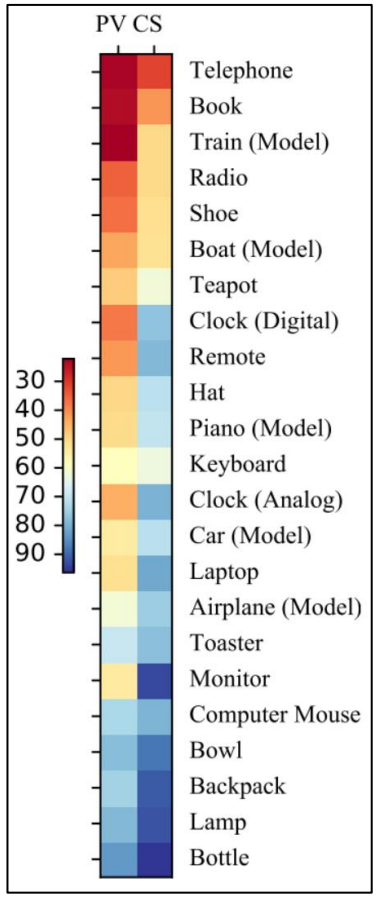
Kobey's Swap Meet in San Diego, CA

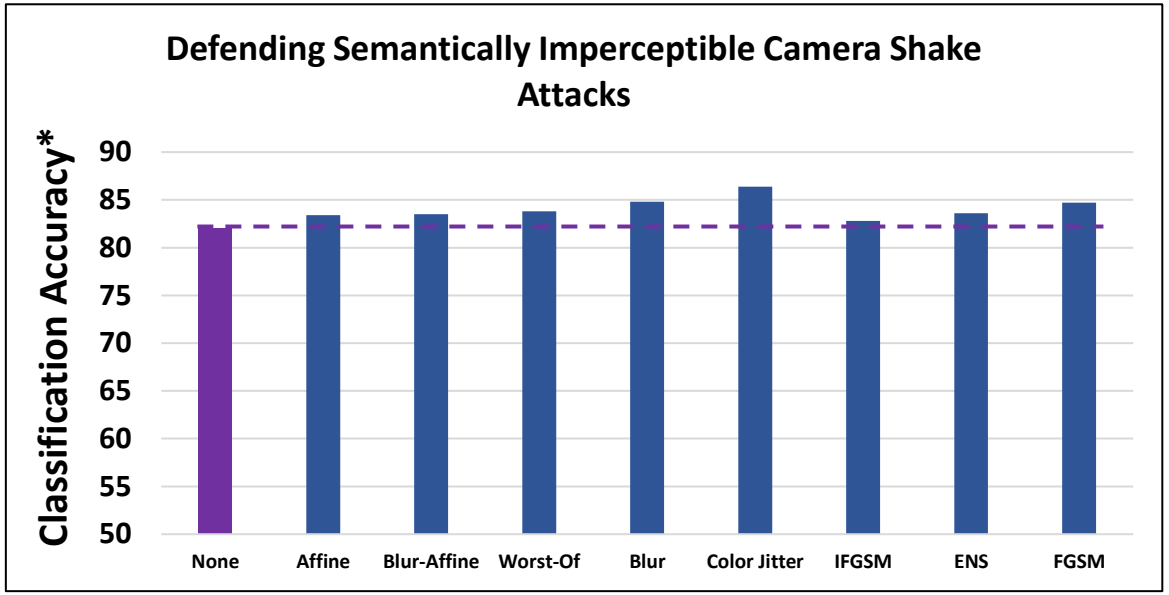
Evaluating Classification Robustness Through Adversarial Attacks

***Leung, B.**, Ho, C. H., Sandstrom, E., Chang, Y., & Vasconcelos, N. (2019). Catastrophic child's play: Easy to perform, hard to defend adversarial attacks.*
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9229-9237.

Robustness of Classifiers on 3D-ODDS







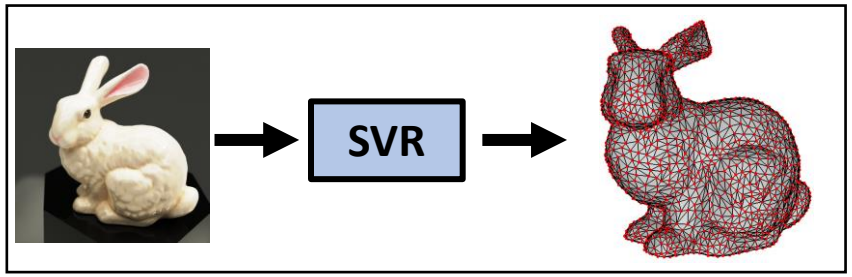
** Tested on different variations of ImageNet pretrained AlexNet, ResNet, & VGG*

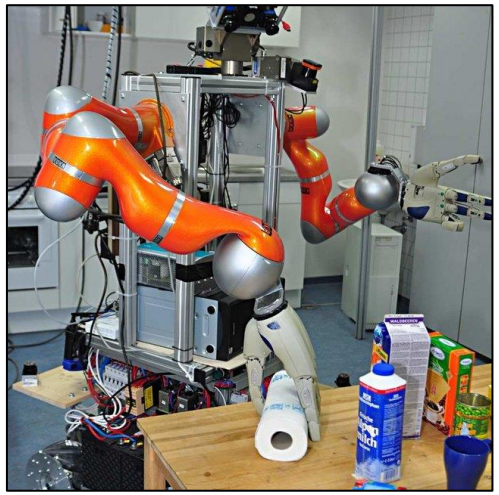
Improving 3D Reconstruction Robustness via Test-Time Adaptation

***Leung, B.**, Ho, C.H., & Vasconcelos, N. (2022). Black-box test-time shape refinement for single view 3d reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Learning with Limited Labelled Data (CVPRW)*

Single View 3D Reconstruction (SVR)

- Goal: Given 2D image of object, recover its 3D shape
- Fundamental problem in 3D vision
  - But very challenging and underconstrained
- Applications:
  - Robotic grasping
  - AR/VR interaction
  - Environment mapping for navigation

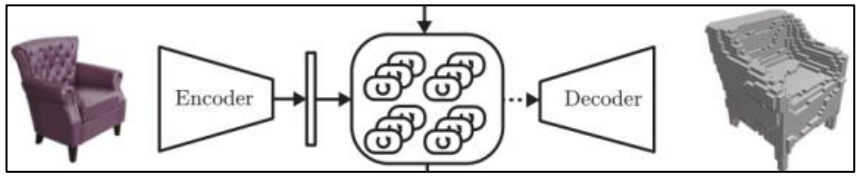




"Total3DUnderstanding", CVPR 20

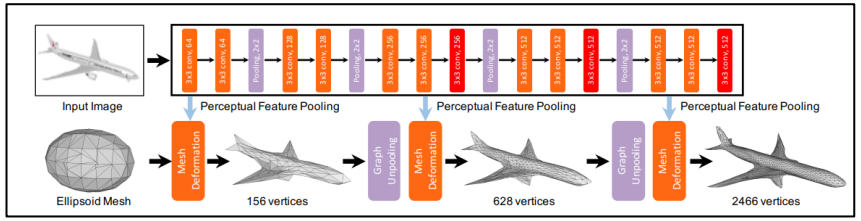
Example SVR Methods

- Encoder-decoder for voxels



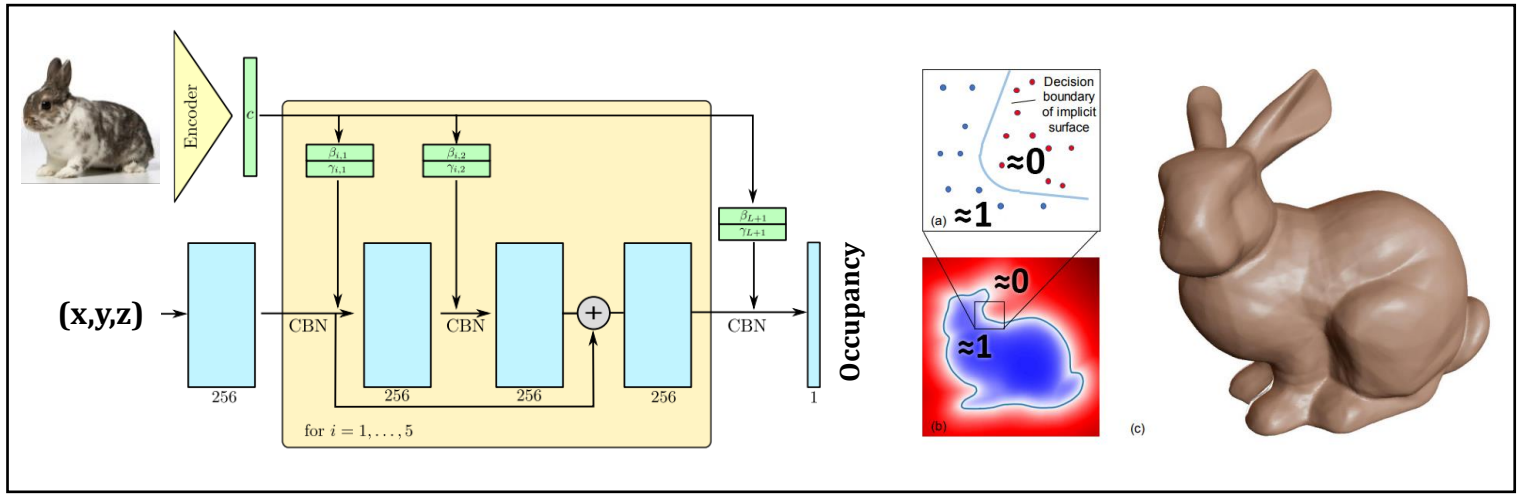
"3D-R2N2", ECCV 16

- Deform Mesh Ellipsoid



"Pixel2Mesh", ECCV 18

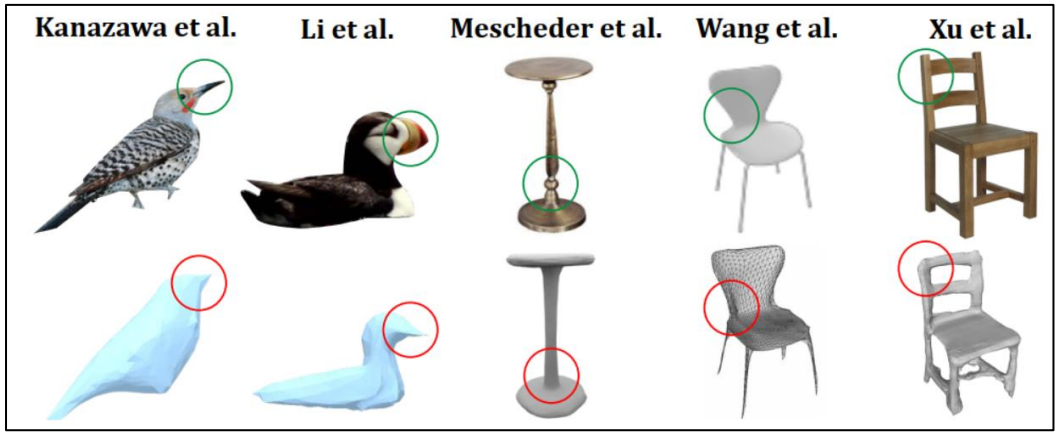
- Learn implicit functional representation



"OccNet", CVPR 19

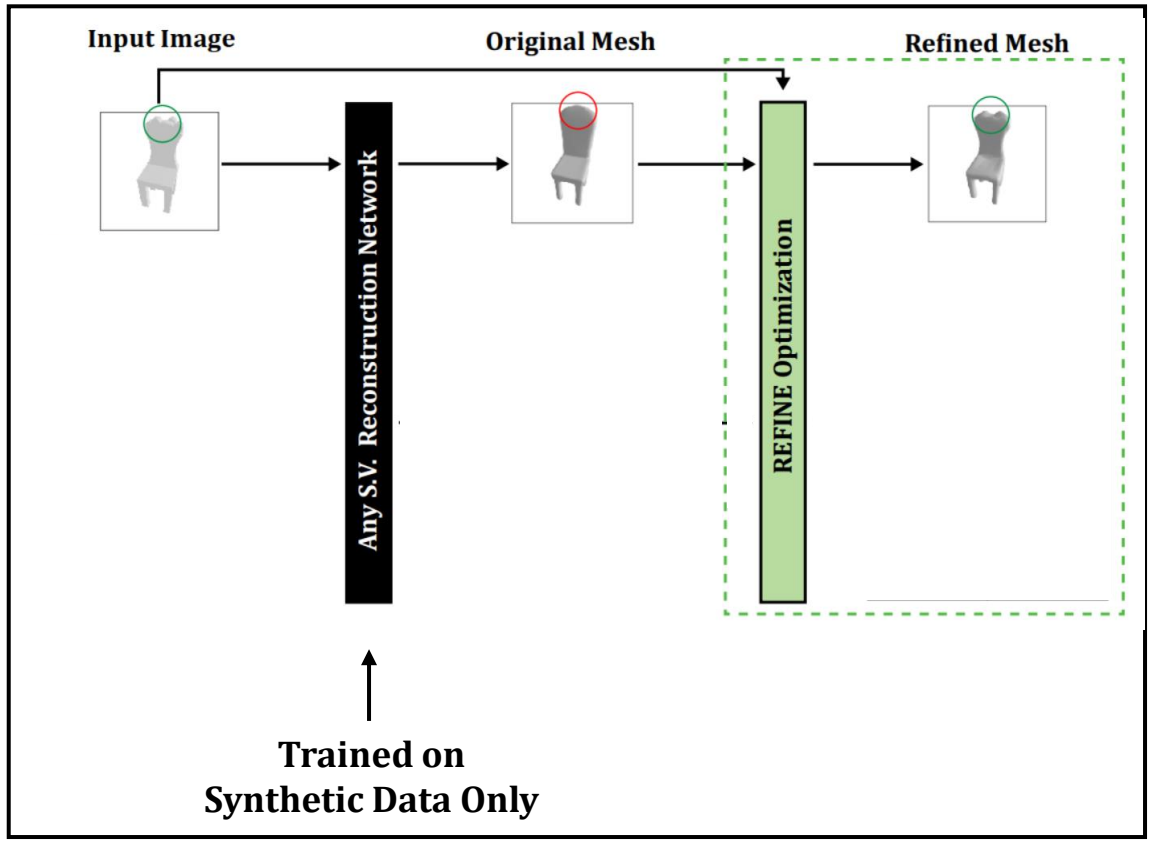
Current Research Issues in Single View Reconstruction

- **Real-world datasets extremely difficult to obtain**
  - Need to rely on synthetic 3D data & their 2D renders, and/or small amount of real data
- **Fine-grained details are hard to capture**
  - Memorization of “mean class shapes” instead of genuine geometric understanding



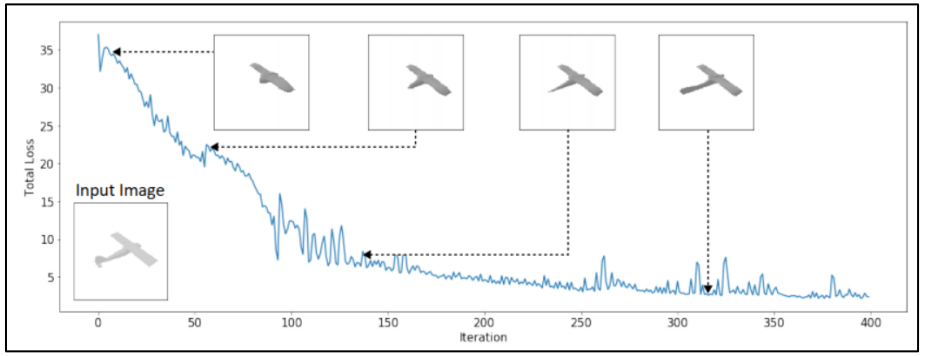
Want to Reduce Domain Gap and Recover Lost Details

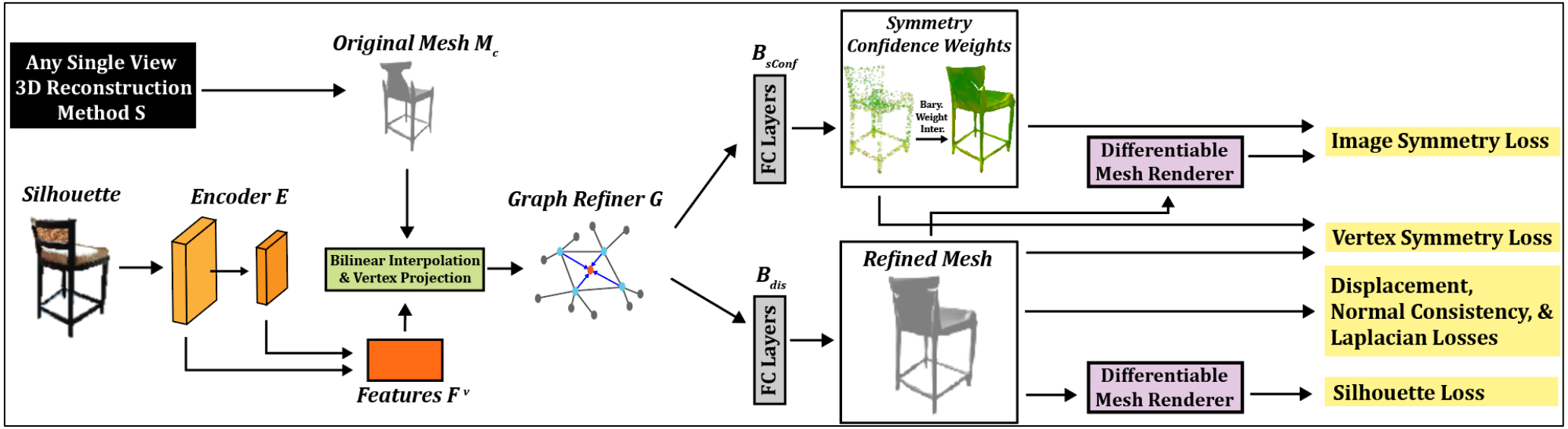
- To do this, we propose a novel **mesh refinement** postprocessing step
- Operates at test-time, treating any base SVR method as a black-box



Given a Bit of Auxiliary Information, Can We Improve 3D Reconstructions?

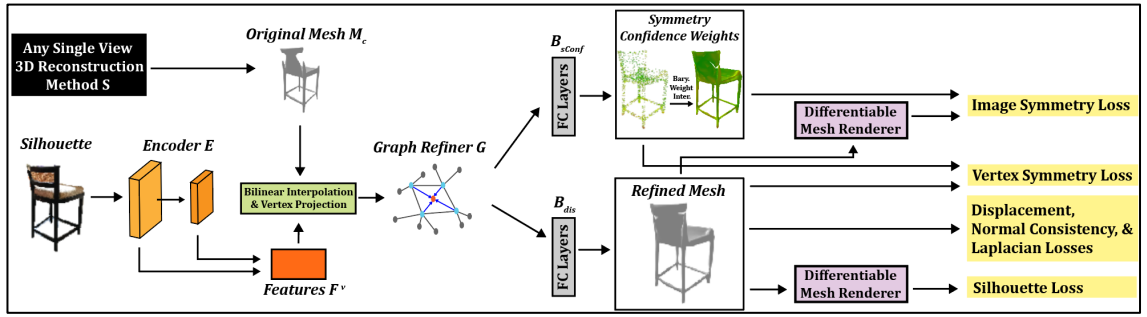
- REFINE neural network learns to displace vertices in a profitable way
  - Exploits silhouette and viewpoint information
  - Goal: have rendered 3D mesh match input image silhouette
  - Several losses provide regularization to prevent degenerate solutions
- Operates at test time, reoptimized **on-the-fly** per mesh





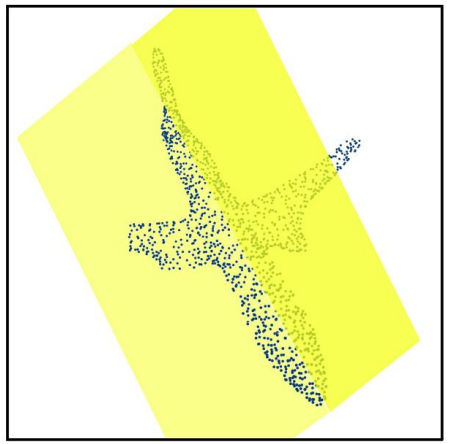
Losses Used in REFINE

1. **Silhouette:** Main information used to improve accuracy
2. **Smoothness:** Laplacian loss & normal consistency loss
3. **Displacement:** Discourages overly large deformations



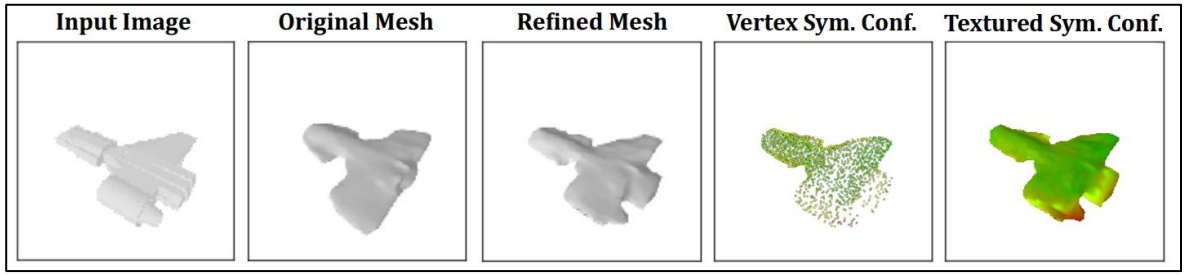
4. Vertex-Based Symmetry:

- Encourages vertices to be close to their symmetric nearest neighbor
- Plane of symmetry used (generally fixed among reconstructions)
- Symmetry confidence weights learned, allowing some asymmetry



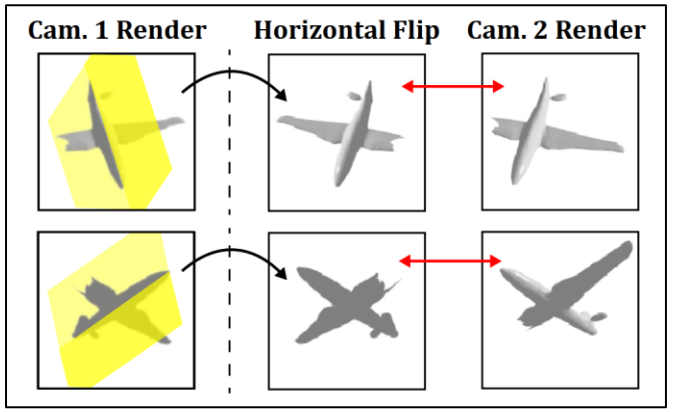
$$L_{Vsym} = \frac{1}{N}\sum_{i=1}^{N}\left[\sigma_i \min_{v_j \in V}\|v_j - Tv_i\|_2^2 + \lambda_{SymB}\ln\left(\frac{1}{\sigma_i}\right)\right]$$

- Avg over verts
- Vert. sym weights
- Sym. nearest neighbor distance
- Vertex asymmetry penalty



5. Image-Based Symmetry:

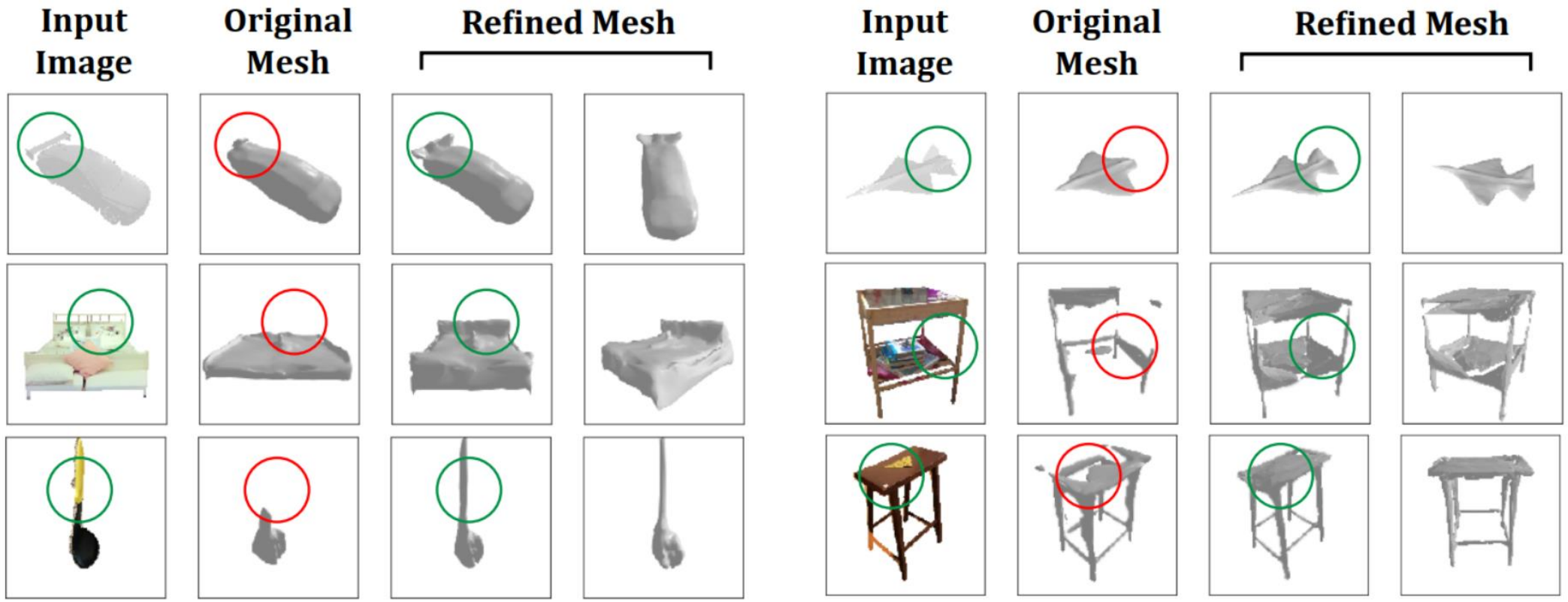
- Encourages render to symmetric
- Symmetry confidence texture by a barycentric interpolation of vertex weights



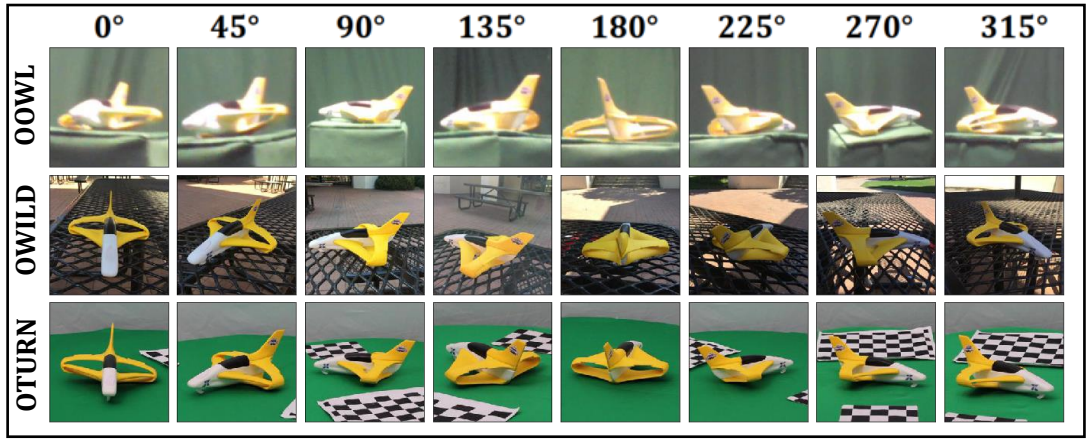
$$L_{Isym} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j,k}\left[\sigma_{jk}\left\|h\left(M_r^{p_i}\right)_{jk} - \left(M_r^{Tp_i}\right)_{jk}\right\|_2^2 + \lambda_{SymB}\ln\left(\frac{1}{\sigma_{jk}}\right)\right]$$

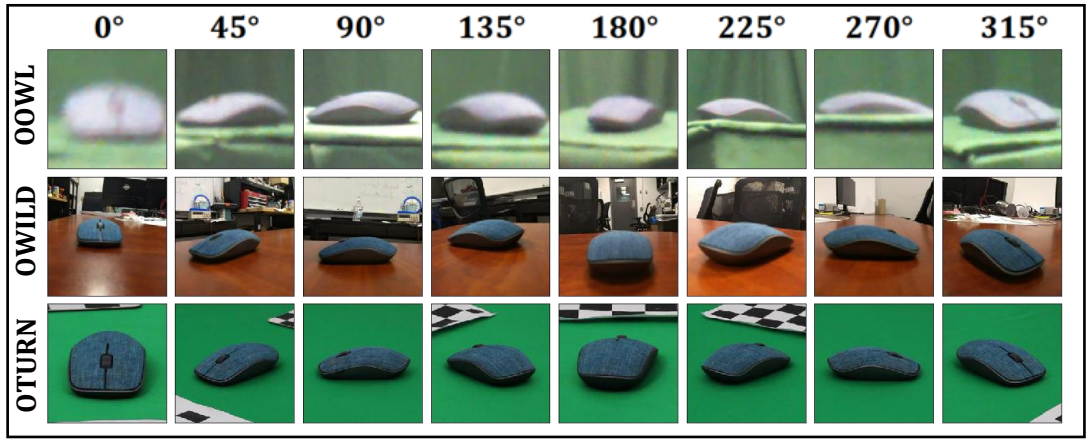
- Avg over cameras
- Pixel-wise sum
- Tex. sym. weights
- Cam render, flipped
- Sym. plane reflected cam render
- Texture asymmetry penalty

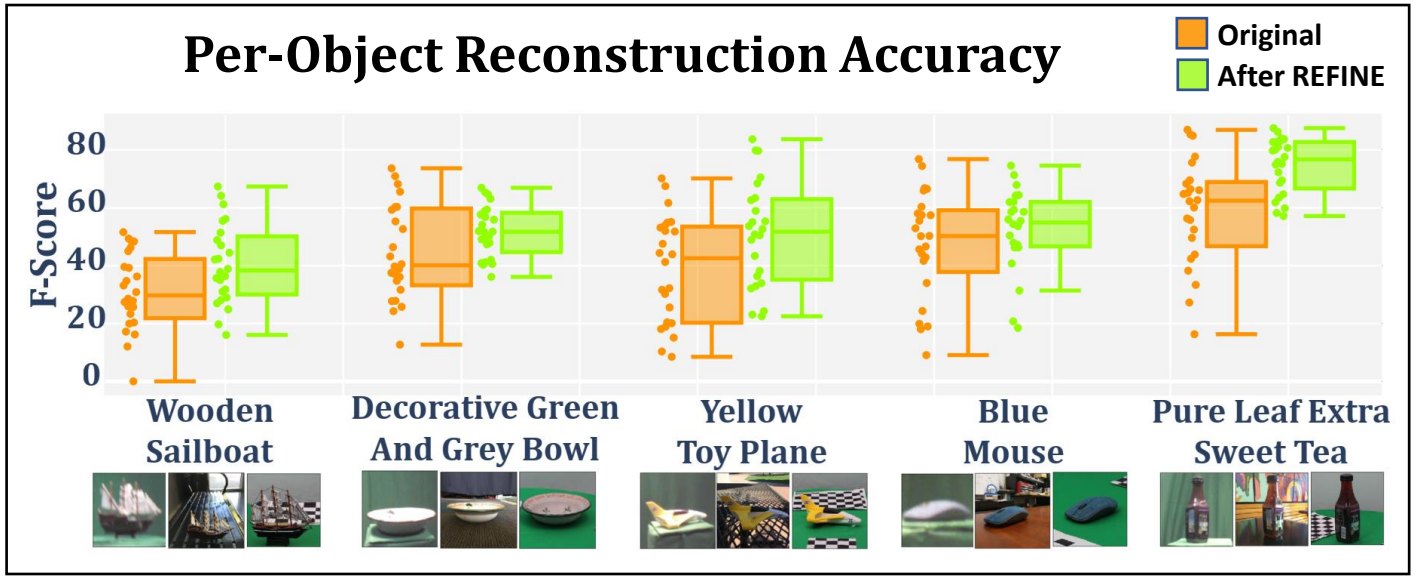
Example Refinements



REFINE on 3D-ODDS





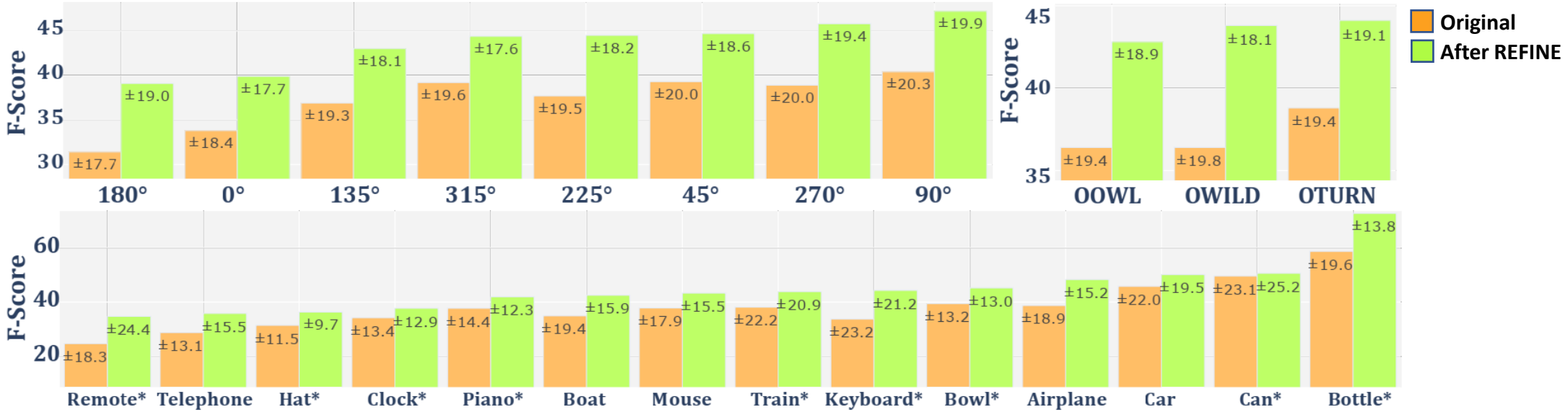


(Original meshes from ShapeNet Trained OccNet)

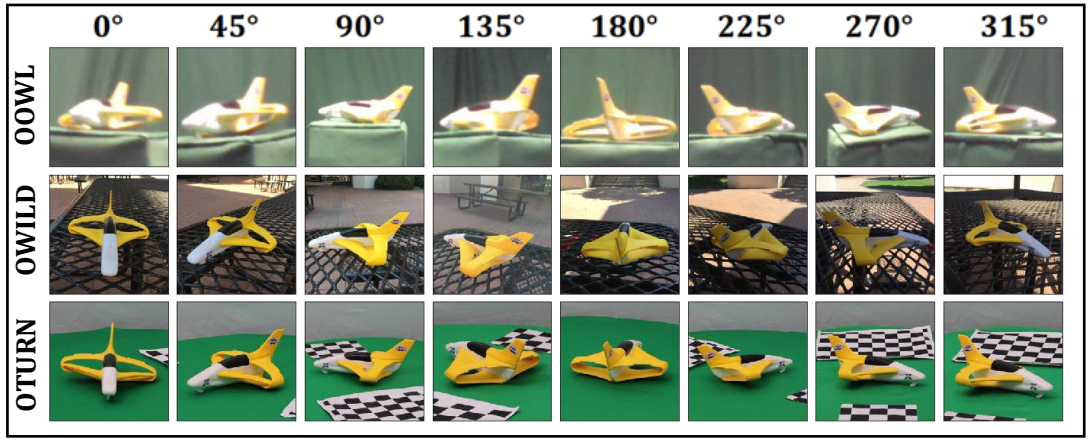
Per-Object-Averaged Accuracy Across 3D-ODDS

	Original	After REFINE
Mean	37.2	44.4
Standard Deviation	16.2	14.3

3D-ODDS Reconstruction Accuracy Across Factors

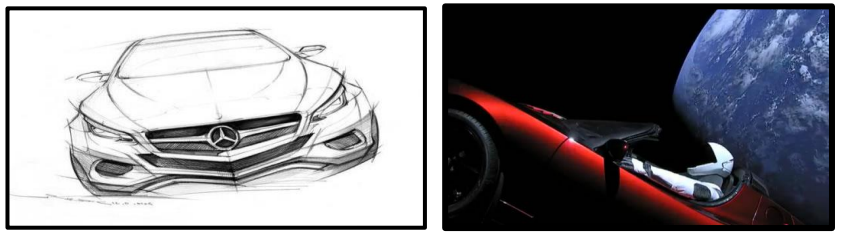


(Original meshes from ShapeNet Trained OccNet)



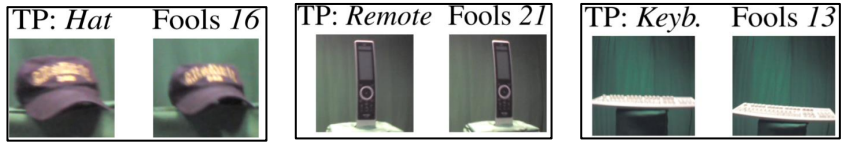
Recap

- **Invariance** is a fundamental, challenging problem for computer vision

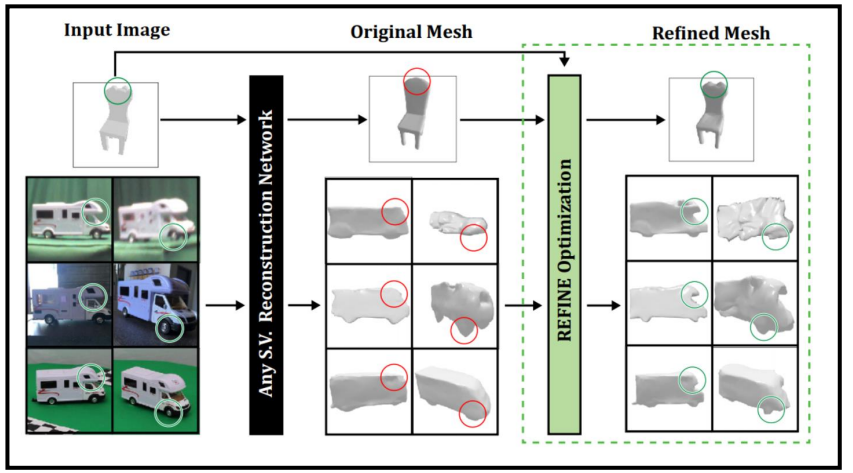


- **3D-ODDS dataset** as a useful testing framework for robustness

- **Natural adversarial robustness attacks** with 3D-ODDS drone images



- **Improving robustness** by refining single view 3D reconstructions



Relevant Publications

- ***Leung, B.**, Ho, C.H., & Vasconcelos, N. (2022). Black-box test-time shape refinement for single view 3d reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Learning with Limited Labelled Data (CVPRW)*

- ***Leung, B.** (2022). Understanding Learned Visual Invariances Through Hierarchical Dataset Design and Collection. MS Thesis.*

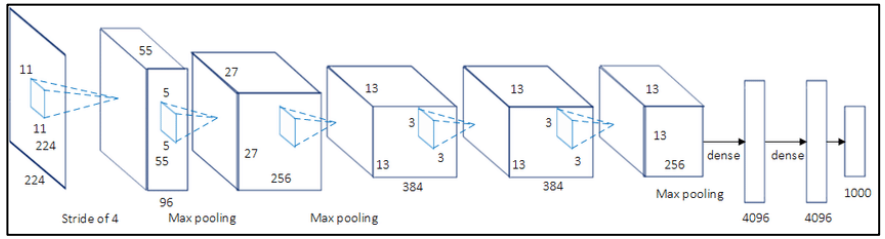
- ***Leung, B.**, Ho, C. H.*, Sandstrom, E., Chang, Y., & Vasconcelos, N. (2019). Catastrophic child's play: Easy to perform, hard to defend adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9229-9237.*

  ** Equal contribution*

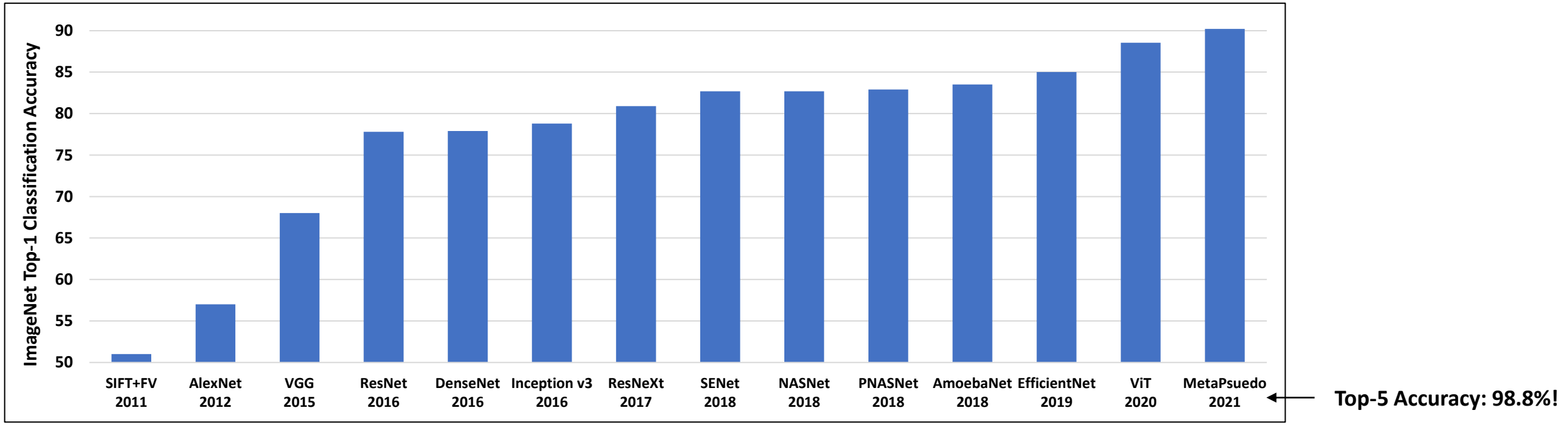
- ***Leung, B.,** Ho, C. H., Persekian, A., Orozco, D., Chang, Y., Sandstrom, E., Liu, B., & Vasconcelos, N. (2019). Oowl500: Overcoming dataset collection bias in the wild. ArXiv:2108.10992*

Additional Materials

- Over the past decade, **Convolutional Neural Networks** (CNNs) have enabled considerable progress towards invariance
  - For many tasks: classification, segmentation, reconstruction, etc



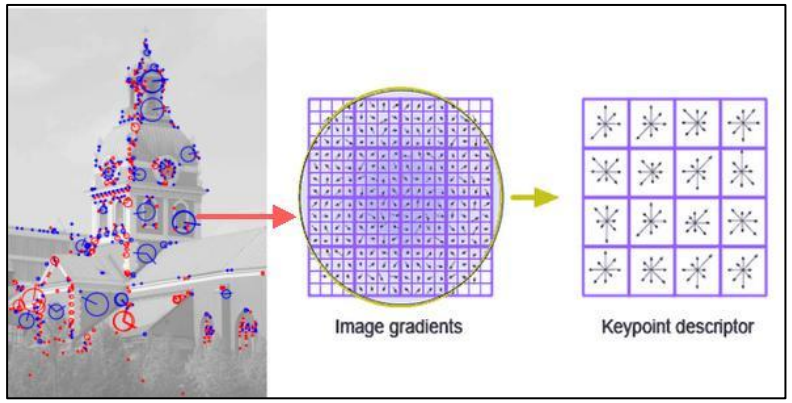
AlexNet, NeurIPS 12



Top-5 Accuracy: 98.8%!

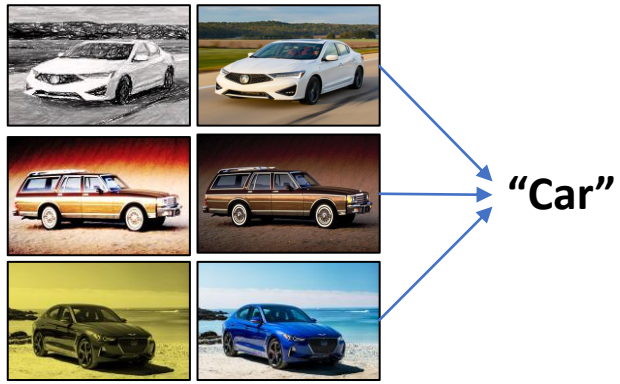
Some Background Context on Invariance

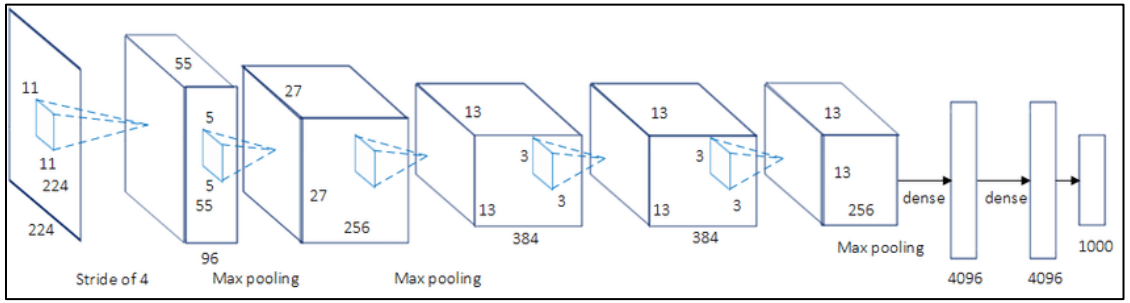
- Classical **handcrafted approaches**, e.g. Harris corner detector or SIFT, only allow for low-level geometric invariances
  - Scale, affine transformation, rotation



Scale-Invariant Feature Transform (SIFT) Features, ICCV 99

- Modern CNNs are **data-driven** and can learn more abstract and complicated invariances
  - Statistical models whose **parameters are learned** through a large **dataset**, to **optimize some defined objective**
  - Downsampling in layers provide **scale invariance**
  - Sliding convolutional filters provide **translation invariance**
  - Other techniques like **data augmentation** also help

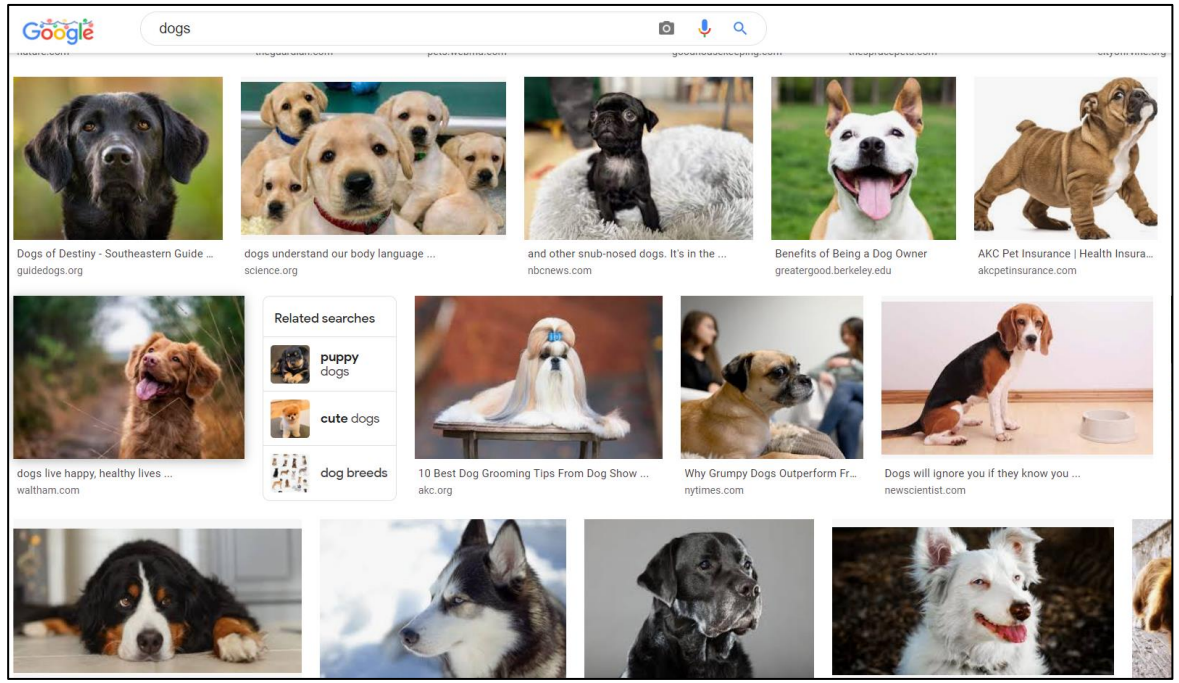


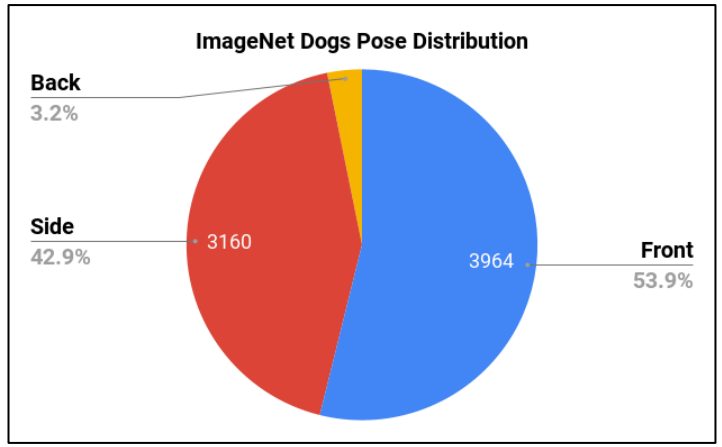


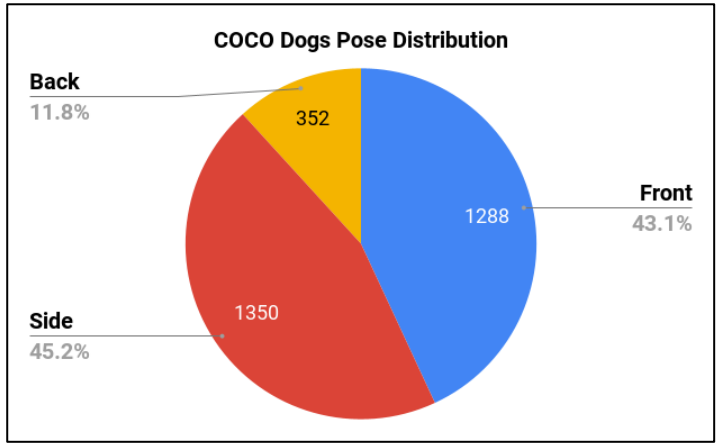
AlexNet, NeurIPS 12

Data-driven CNNs are a double-edged sword

- In many cases, **drastically increased performance**
- Can be **unstable and unintuitive**
  - Millions of parameters approximately optimized through gradient descent
- Subject to **dataset bias**
  - Viewpoint bias, lighting bias, camera quality bias, etc

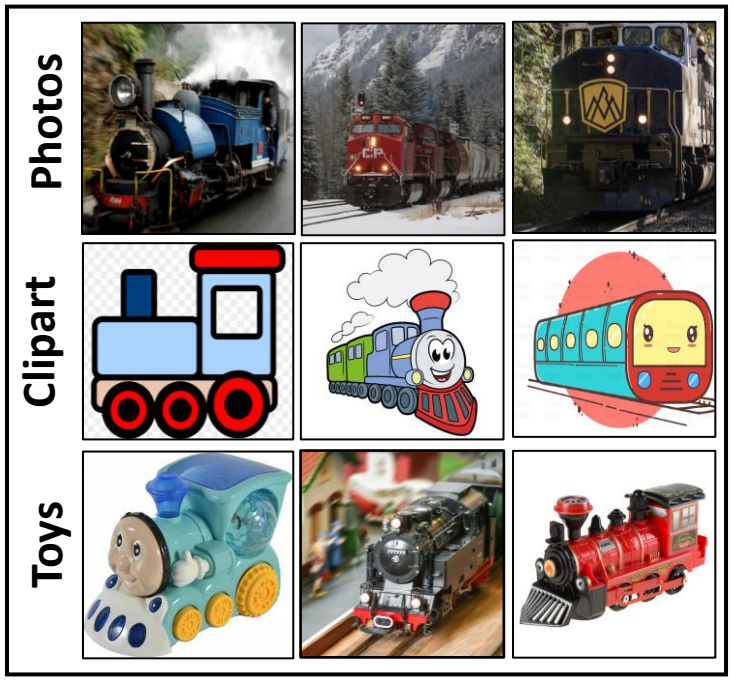






Can we just curate a web-scraped image dataset?

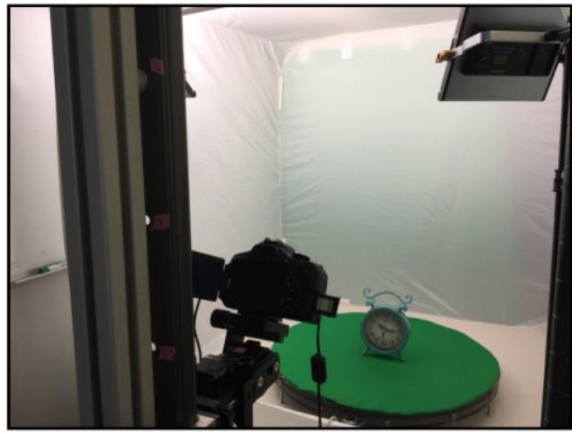
Web-Scraped Data Curation

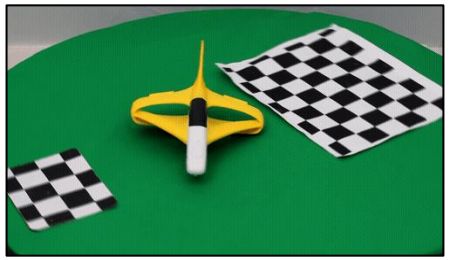


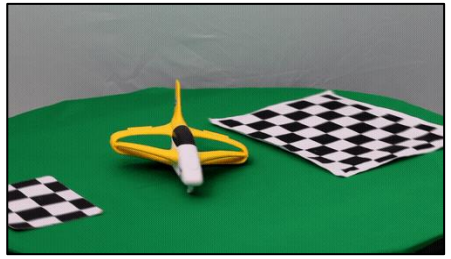
- **Simple & cost effective; diverse**
- **Individual objects scattered & seen inconsistently**, leading to measurement ambiguity
- Subject to **uneven viewpoint coverage**
- **Only 2D**, in general no access to 3D information
- May have **within-domain inconsistency**; different lighting, camera quality, etc within a domain

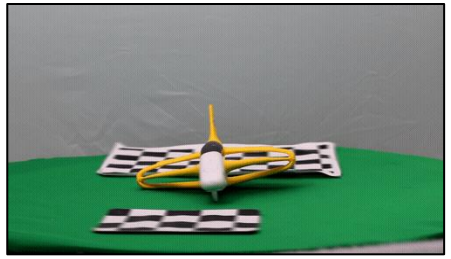
OTURN Domain: Dataset Design & Collection

- Collected in the lab, using a turntable and DSLR camera setup
- Blank background, professional-grade lighting
- 3D meshes obtained using structure-from-motion software

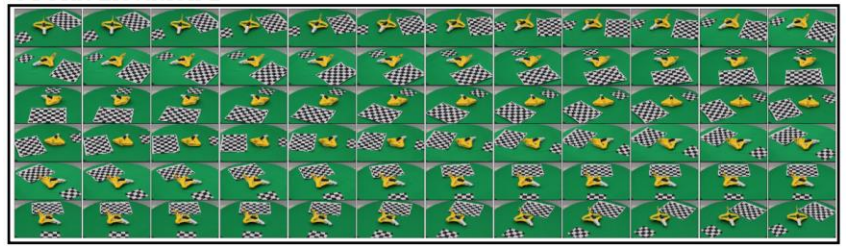




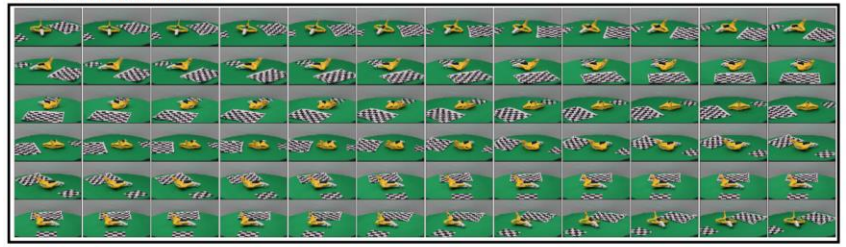




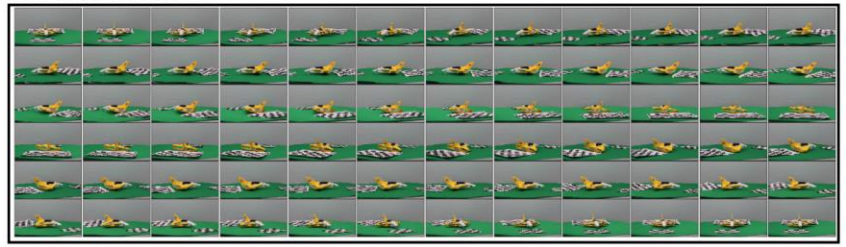
OTURN Elevation 1

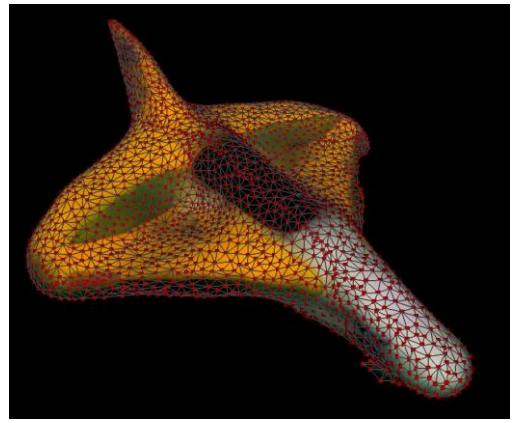


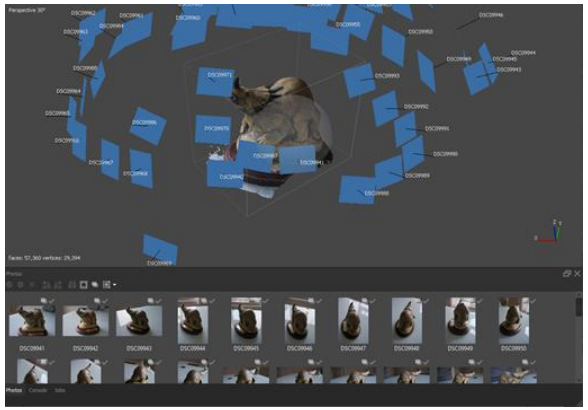
OTURN Elevation 2



OTURN Elevation 3

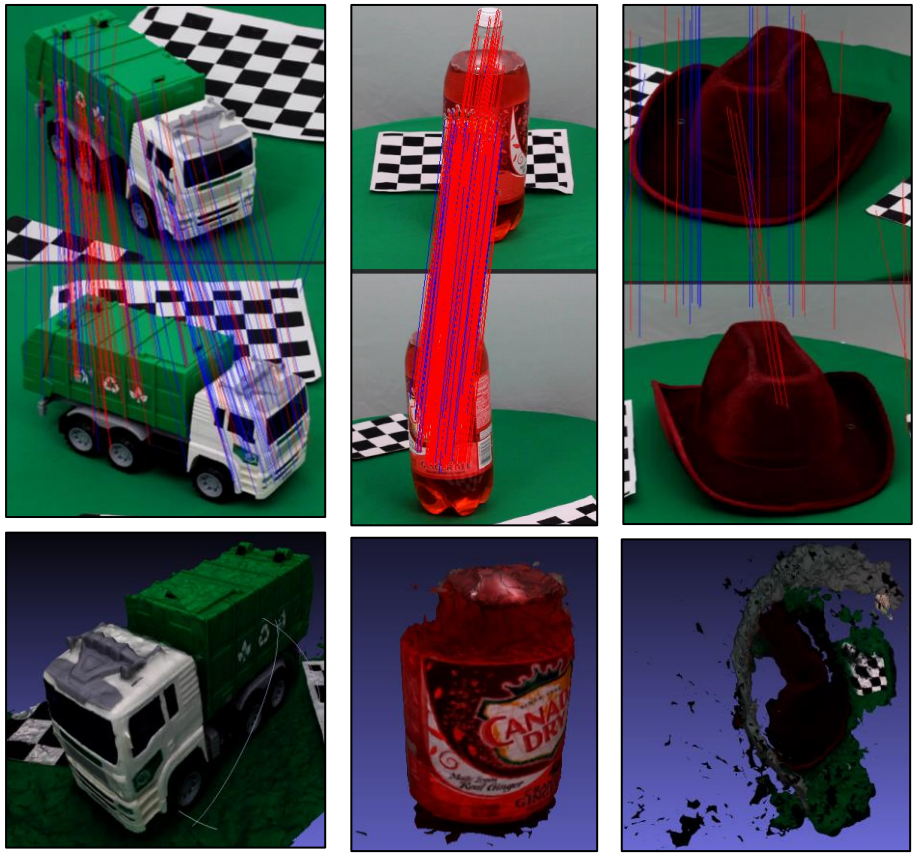


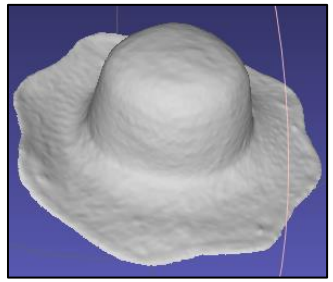




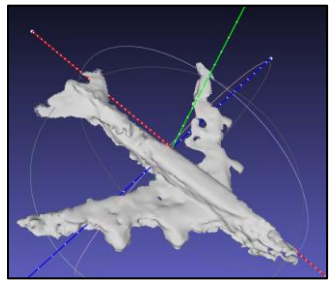
Noise is a natural consequence of real-world 3D data collection

- Primarily comes from mismatched correspondences, due to absence of texture or excessive reflectance
- We only use high/medium quality OTURN meshes
  - 101 high quality
  - 198 medium quality
  - 32 low quality

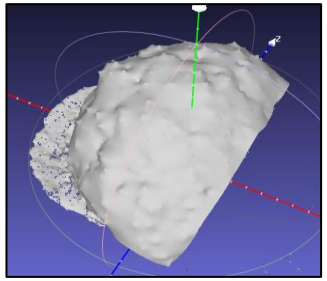




High Quality



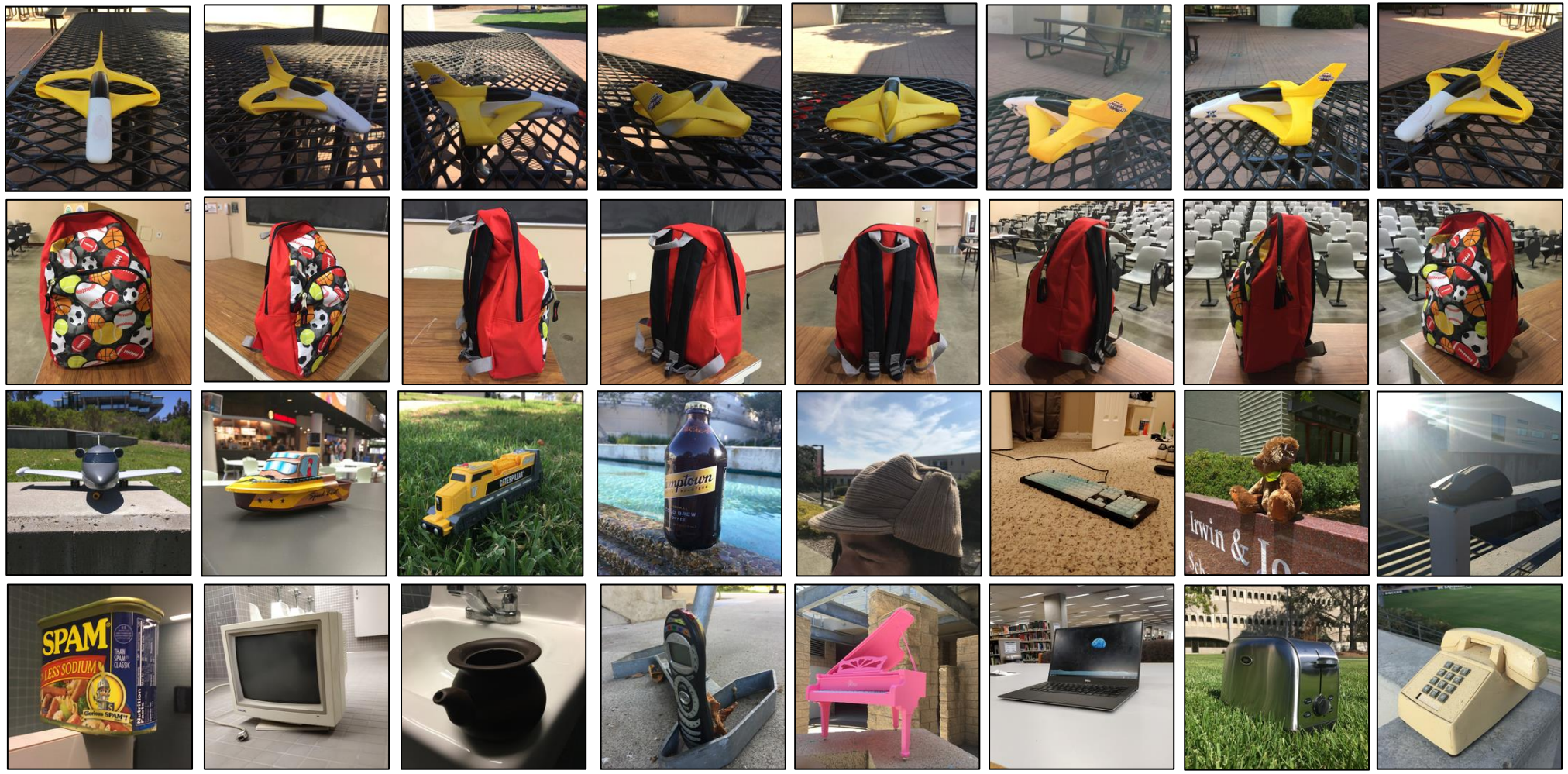
Medium Quality



Low Quality

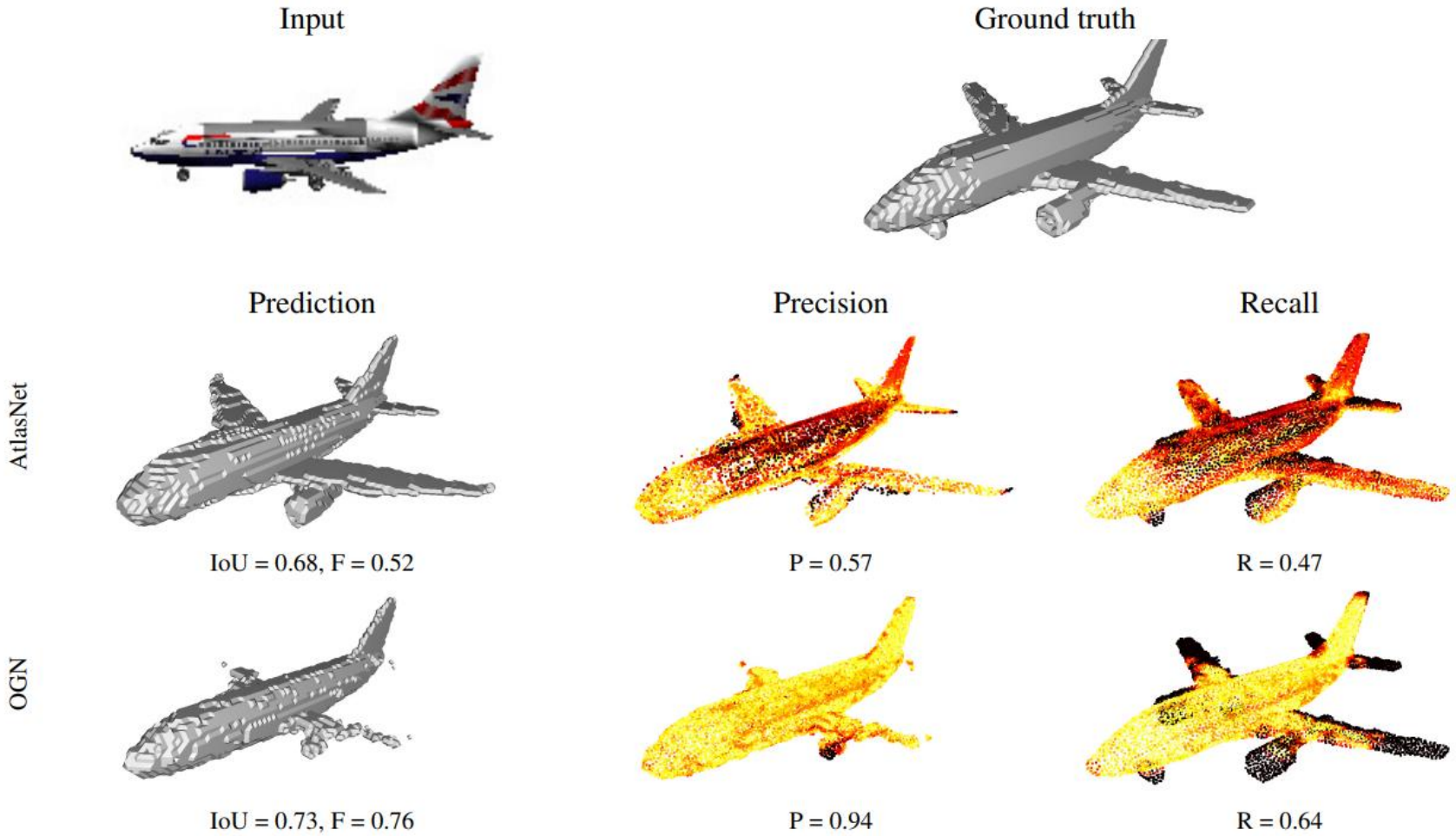
OWILD Domain: Dataset Design & Collection

- Collected around the UCSD campus, using smartphones
- Diverse indoor/outdoor scenes

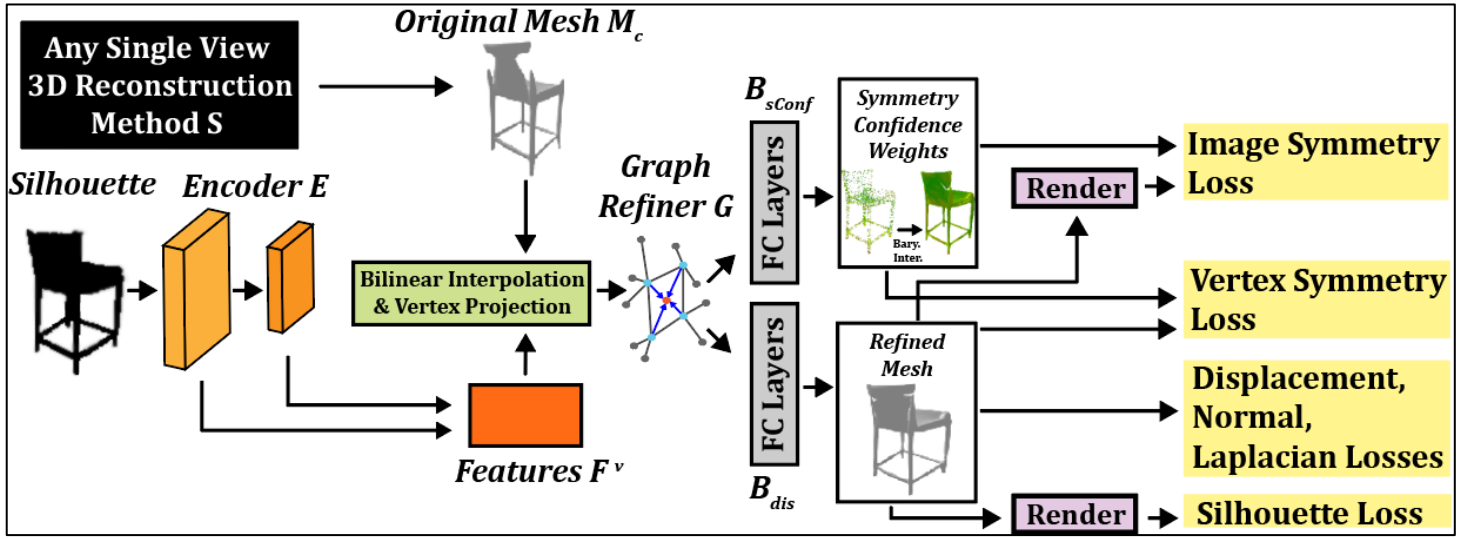


F-Score

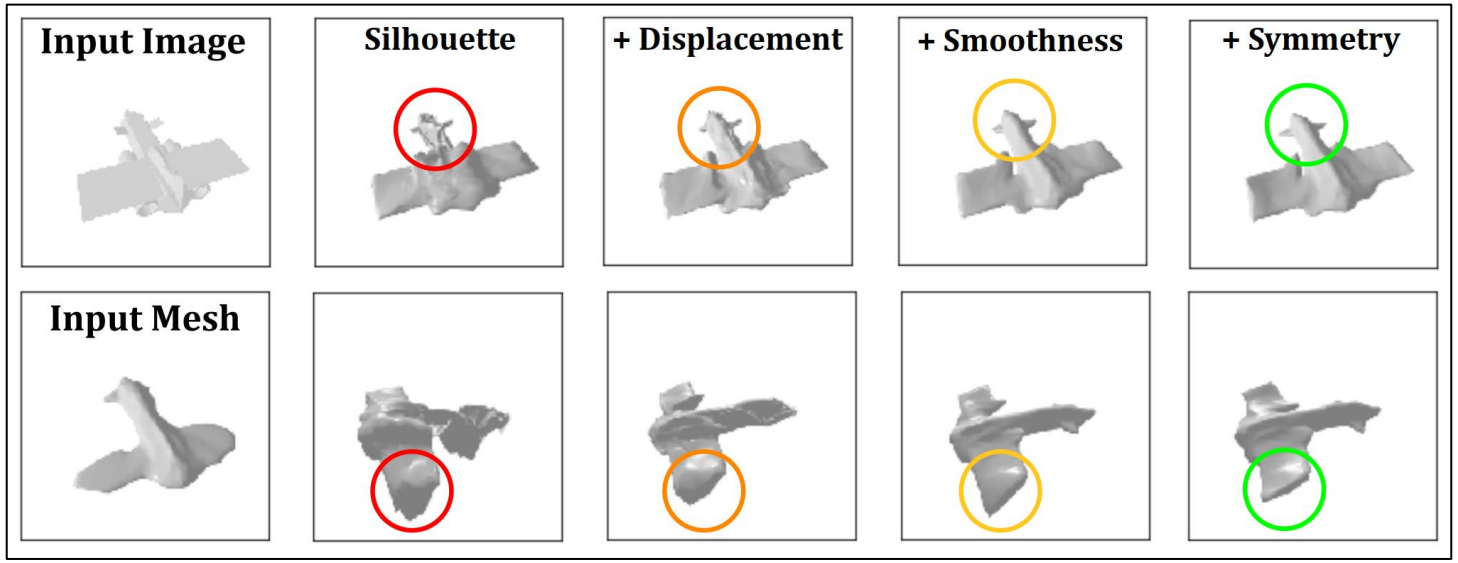
- Harmonic mean between precision and recall
- **Precision (accuracy of reconstruction)**: percentage of reconstructed points that lie within a certain distance to the ground truth
- **Recall (completeness of reconstruction):** percentage of points on the ground truth that lie within a certain distance to the reconstruction



$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

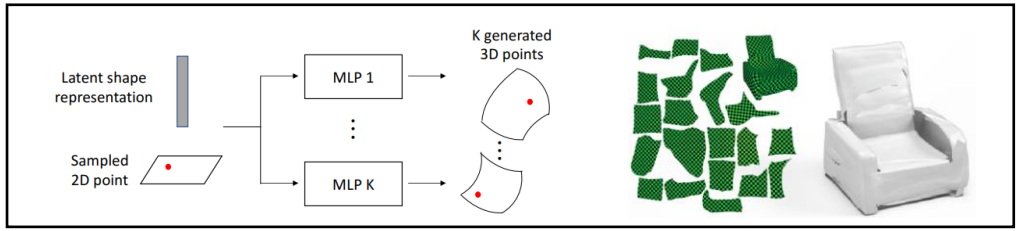


Configuration	EMD↓	CD-l_2 ↓	F-Score↑	Vol. IoU↑	2D IoU
OccNet [34]	4.3	34.0	80	33	69
L_{Sil}	12.2	154.8	51	16	87
$L_{Sil,Dis,Nc,Lp}$	3.7	26.2	80	31	85
$L_{Sil,Dis,Nc,Lp,Vsym}$	3.7	25.8	81	32	86
$\boldsymbol{L_{total}}$	**3.3**	**22.5**	**84**	**35**	85
E & G removed, L_{total}	3.5	24.5	82	33	87
E removed, L_{total}	3.4	24.1	82	34	87
E rand. init, L_{total}	3.4	23.1	83	35	85
OccNet* [34]	11.0	123.3	48	10	53
$L_{total}, \lambda_{SymB} = 1.0$*	8.9	89.1	52	10	72
$\boldsymbol{L_{total}}$*	**7.8**	**85.9**	**55**	**12**	76

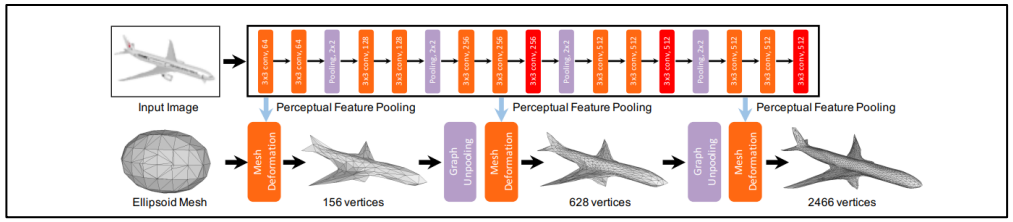


REFINE Efficacy Over Different SVR Methods

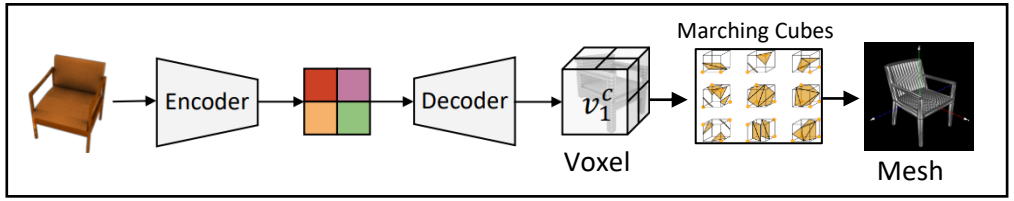
"AtlasNet" CVPR 18



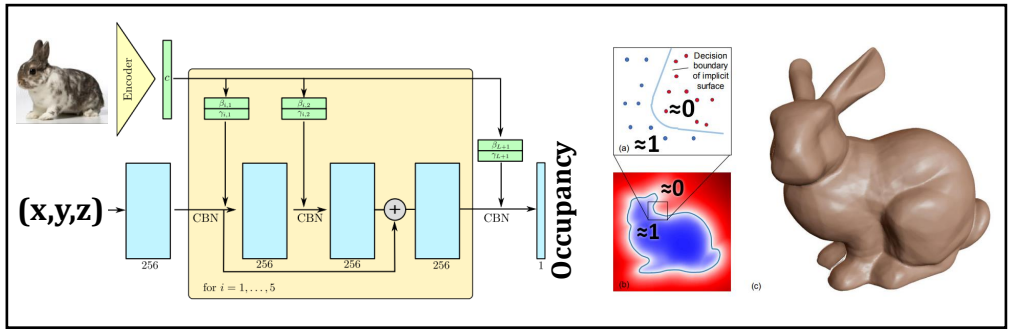
"Pixel2Mesh" ECCV 18

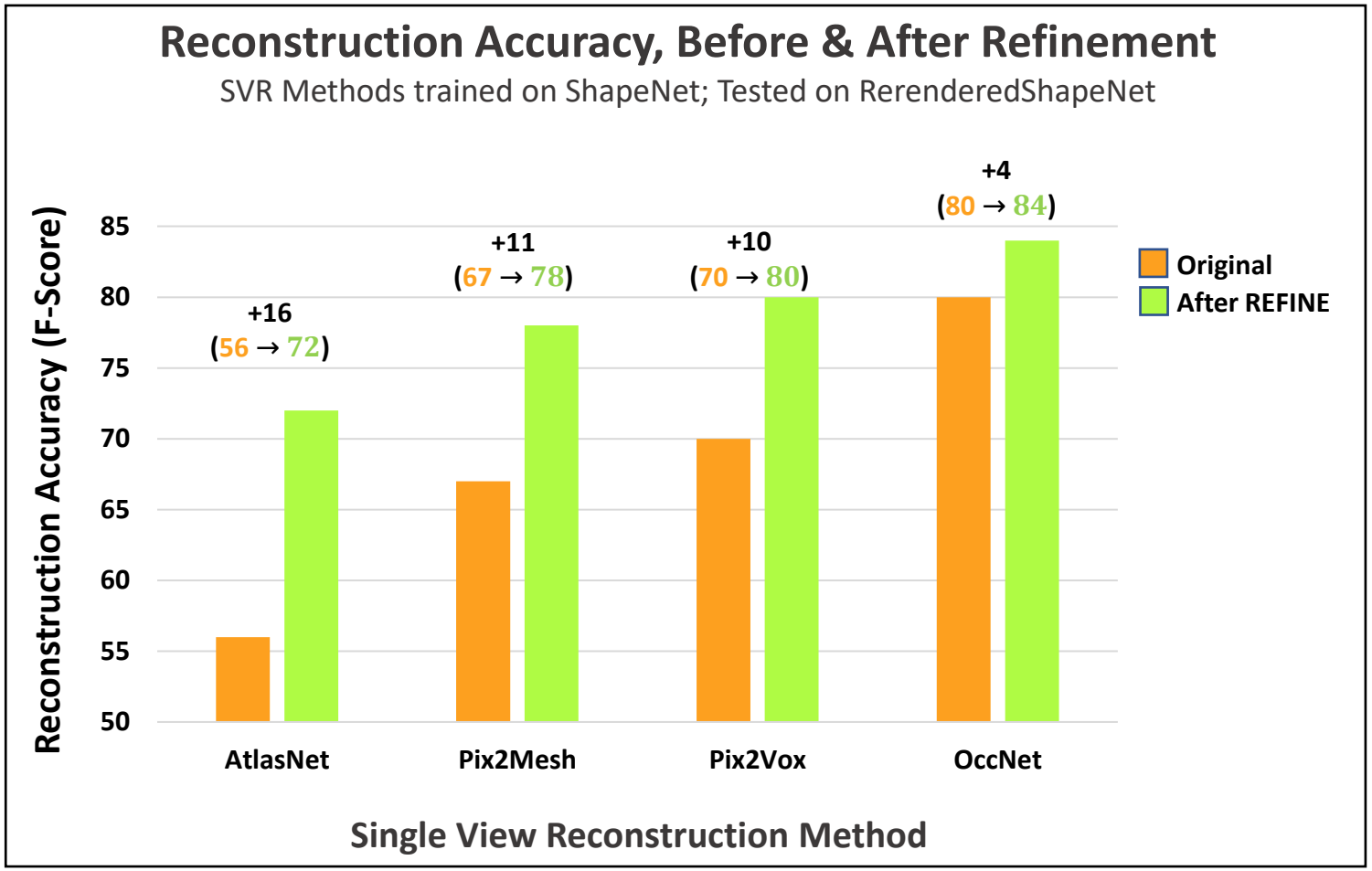


"Pix2Vox" IJCV 20

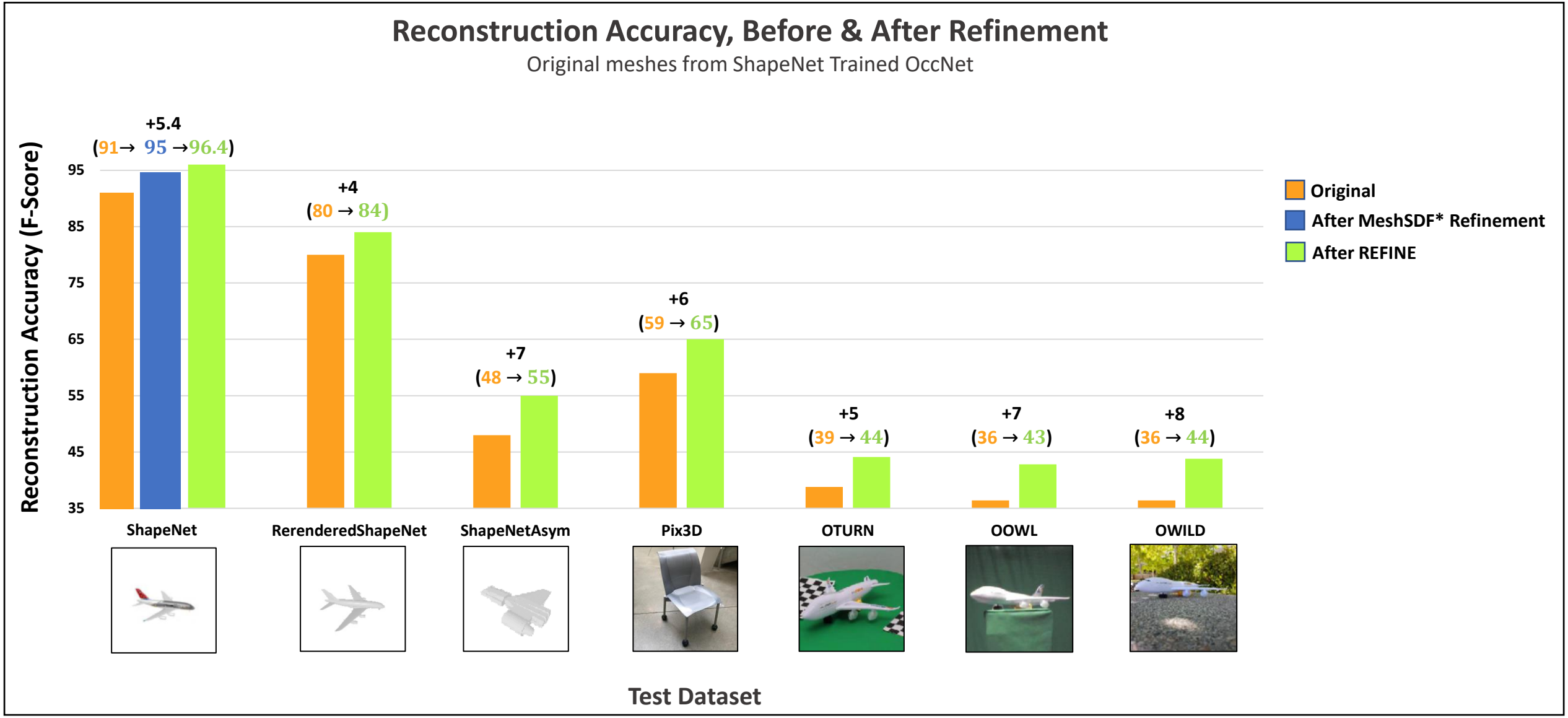


"OccNet" CVPR 19

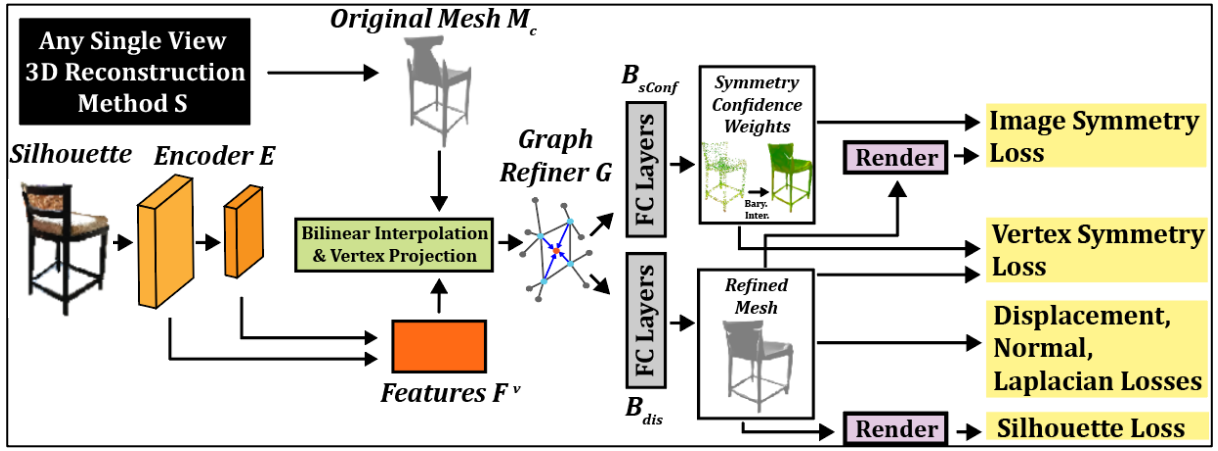


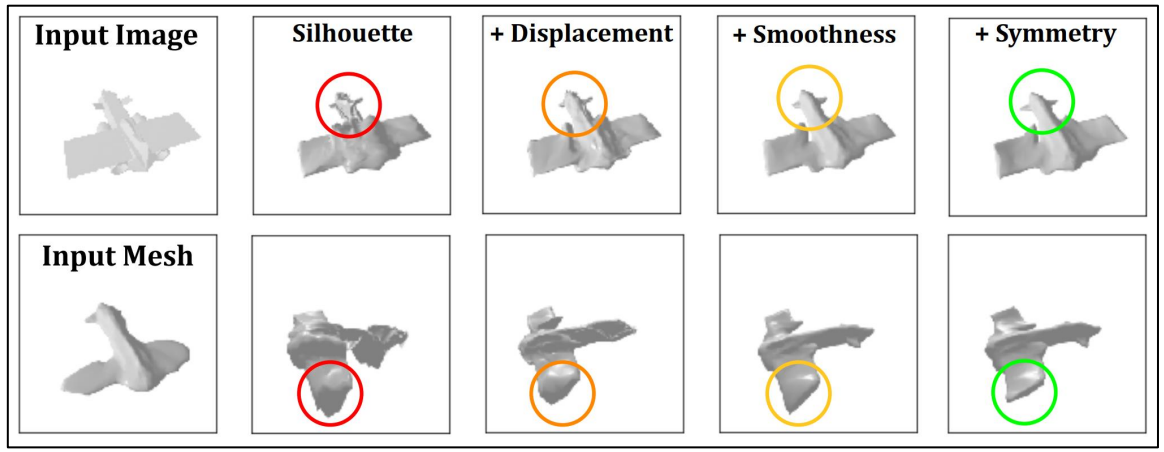


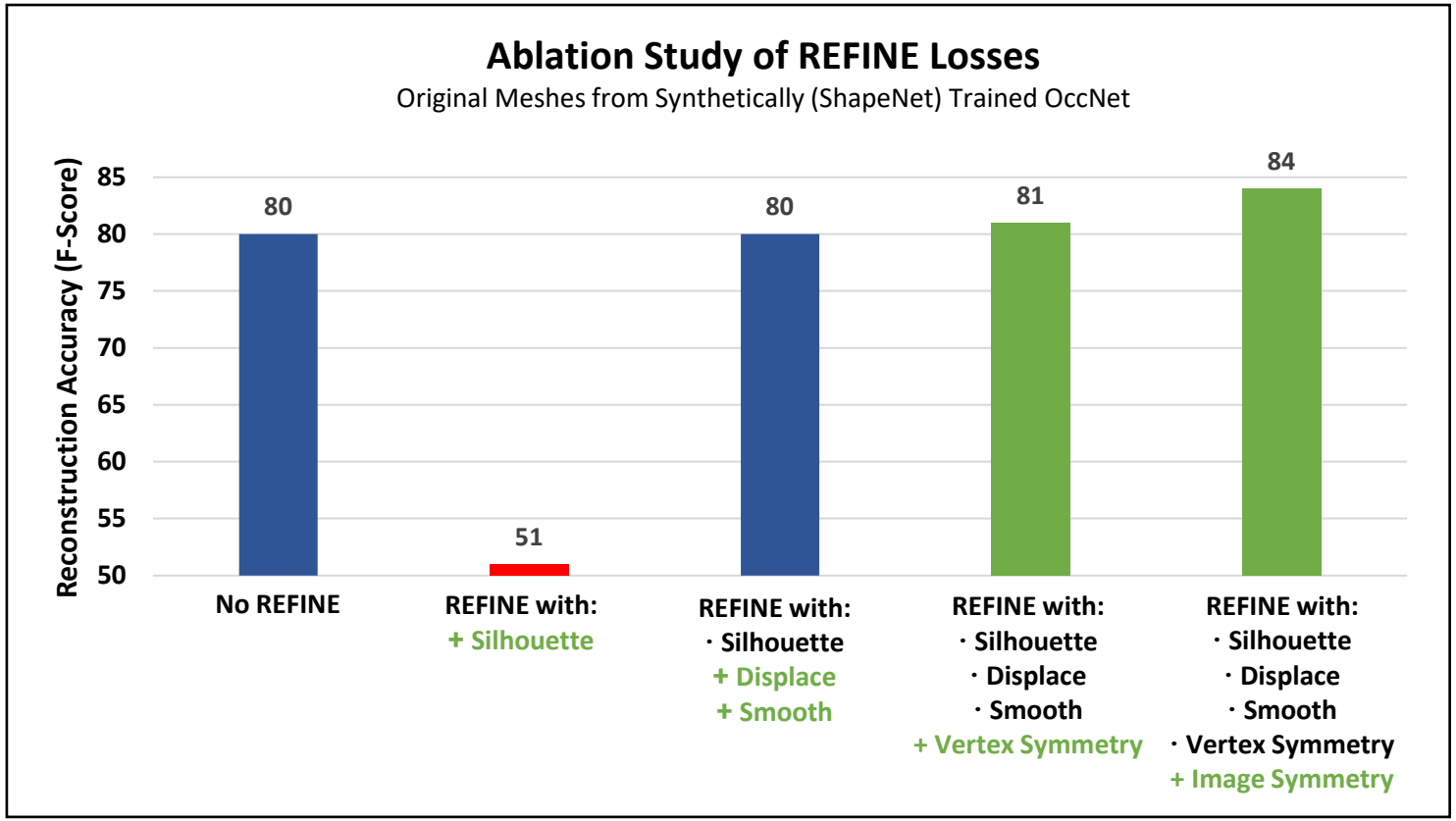
REFINE Efficacy Over Different Datasets

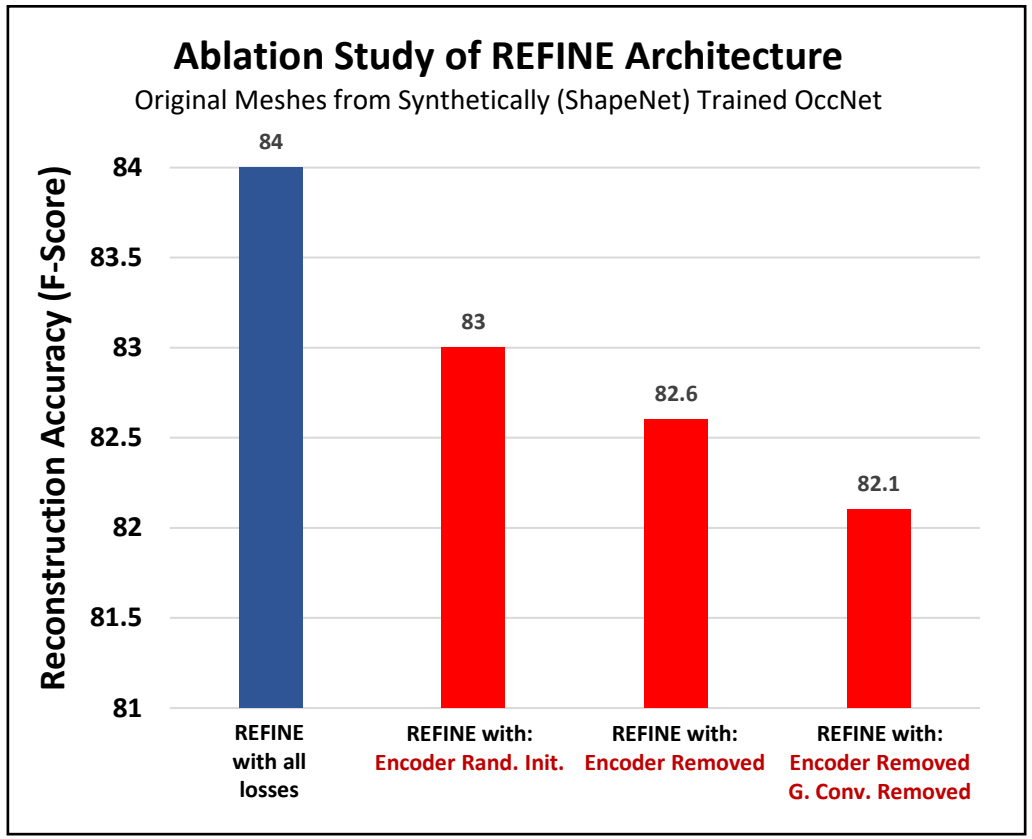


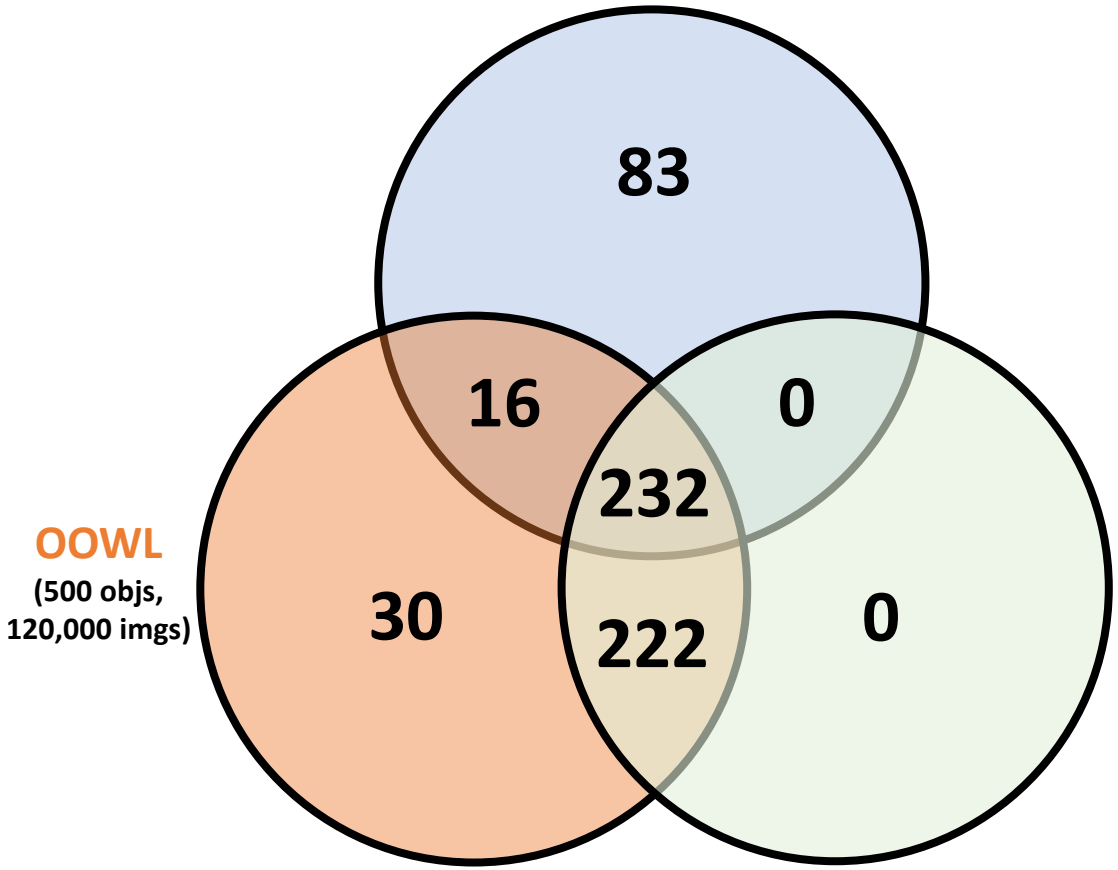
** **White-box** refinement applicable only to implicit function methods, NeurIPS 20.*


Any Single View 3D Reconstruction Method S
Original Mesh M_c
Silhouette
Encoder E
Graph Refiner G
Bilinear Interpolation & Vertex Projection
Features F^v
B_{sConf}
FC Layers
Symmetry Confidence Weights
Bary. Inter.
Render
Image Symmetry Loss
Vertex Symmetry Loss
FC Layers
Refined Mesh
Displacement, Normal, Laplacian Losses
B_{dis}
Render
Silhouette Loss


Input Image
Silhouette
+ Displacement
+ Smoothness
+ Symmetry
Input Mesh


Ablation Study of REFINE Losses
Original Meshes from Synthetically (ShapeNet) Trained OccNet
Reconstruction Accuracy (F-Score)
85
80
75
70
65
60
55
50
80
51
80
81
84
No REFINE
REFINE with: + Silhouette
REFINE with: · Silhouette + Displace + Smooth
REFINE with: · Silhouette · Displace · Smooth + Vertex Symmetry
REFINE with: · Silhouette · Displace · Smooth · Vertex Symmetry + Image Symmetry


Ablation Study of REFINE Architecture
Original Meshes from Synthetically (ShapeNet) Trained OccNet
Reconstruction Accuracy (F-Score)
84
83.5
83
82.5
82
81.5
81
84
83
82.6
82.1
REFINE with all losses
REFINE with: Encoder Rand. Init.
REFINE with: Encoder Removed
REFINE with: Encoder Removed G. Conv. Removed


OTURN
(331 objs,
71,496 imgs,
331 3D-meshes)
83
16
0
232
OOWL
(500 objs,
120,000 imgs)
30
222
0
OWILD
(454 objs,
3632 imgs)

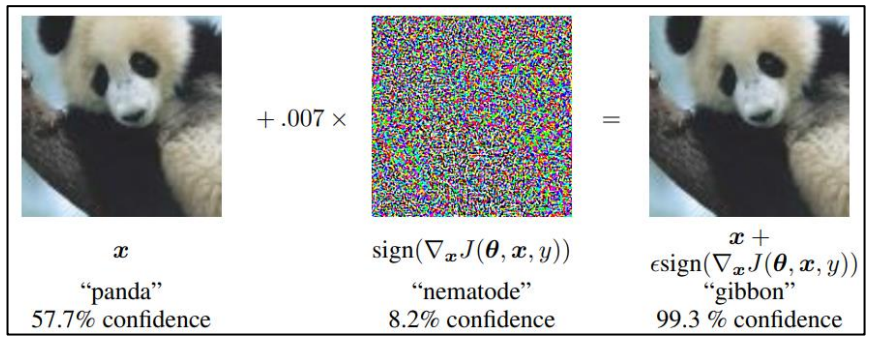

OOWL
OWILD
OTURN

Evaluating Classification Robustness Through Adversarial Attacks

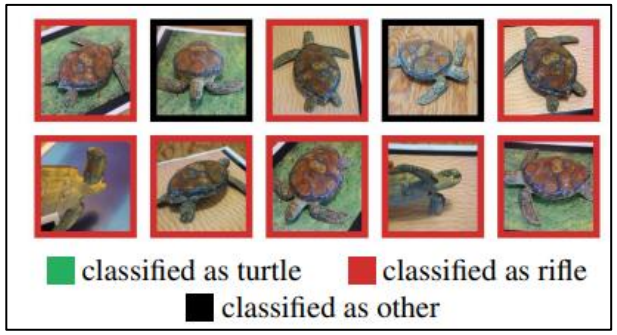
Ho, C. H., **Leung, B.***, Sandstrom, E., Chang, Y., & Vasconcelos, N. (2019). Catastrophic child's play: Easy to perform, hard to defend adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9229-9237.*

Adversarial Attacks on CNNs

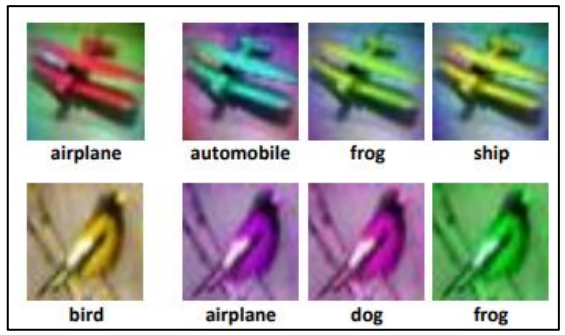
- Adversarial attacks are images **minimally modified** such that CNNs **fail dramatically**
- Of great concern for many safety-critical applications
- Defenses usually involve retraining networks with attacks
- **Arms-race/coevolution** between attacks and defenses



Explaining and Harnessing Adversarial Examples, ICLR 15

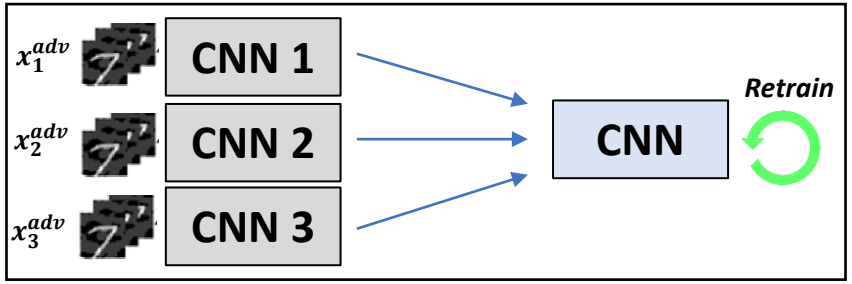


Synthesizing Robust Adversarial Examples, ICML 18

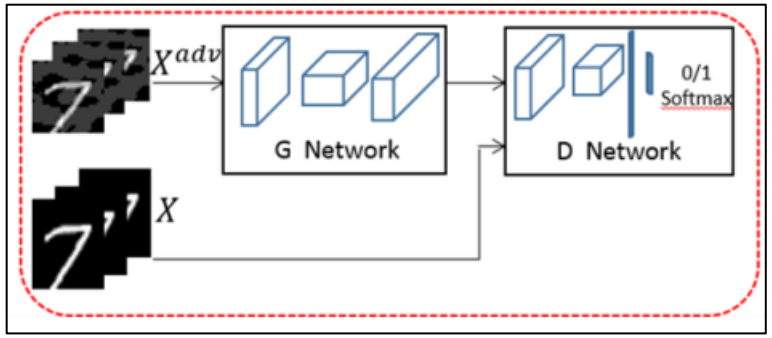


Semantic Adversarial Examples, CVPR 18 Workshop

Adversarial Attacks



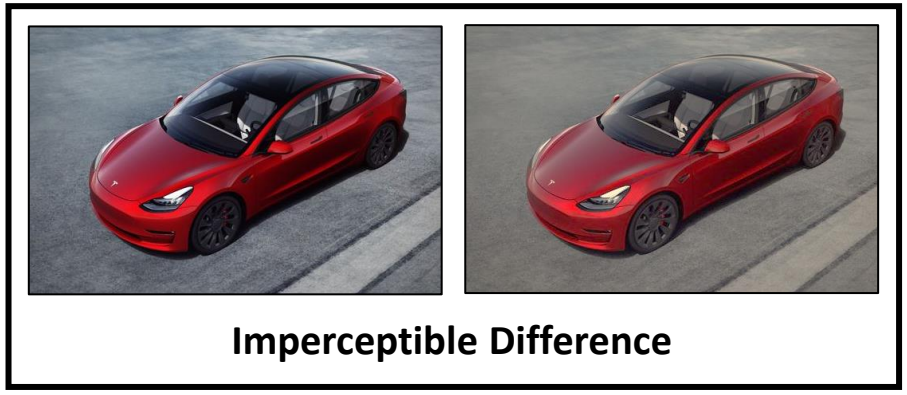
Ensemble adversarial training: Attacks and defenses, ICLR 18



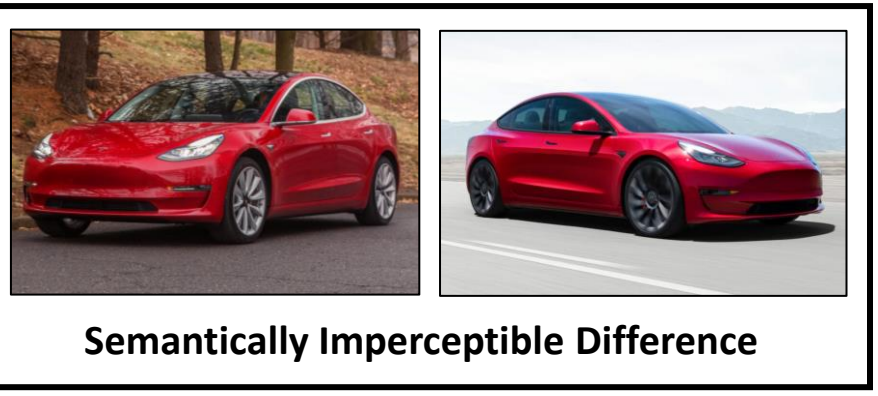
APE-GAN, ICASSP 2019

Adversarial Defenses

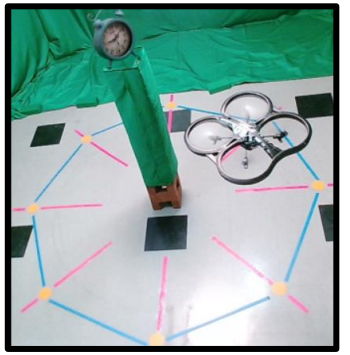
- Want insight into **computer vision vs human vision** robustness
- Usually, “minimally modified” enforced with arbitrarily small thresholds to make attacks imperceptible
- In this work, we instead use **human-based perception**
  - **Perception categories**: Imperceptible & Semantically Imperceptible
  - **Perturbation types**: Camera shake and viewpoint variation, using the **OOWL dataset**
- These attacks **leverage the limitations of computer vision**
  - Easily produced by people but hard to defend against, since they cannot be replicated by computers

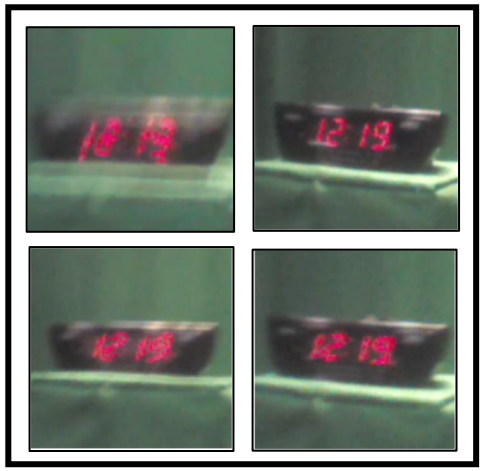


Imperceptible Difference



Semantically Imperceptible Difference



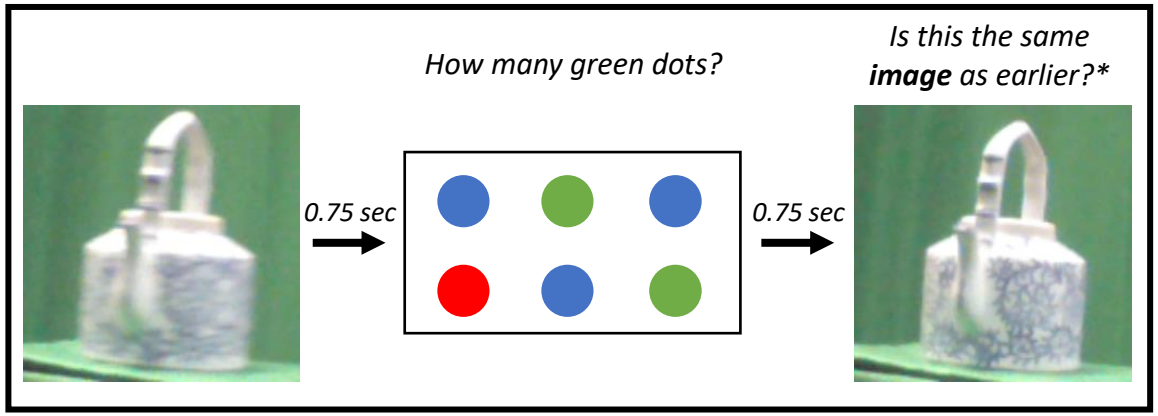


Camera Shake Attack

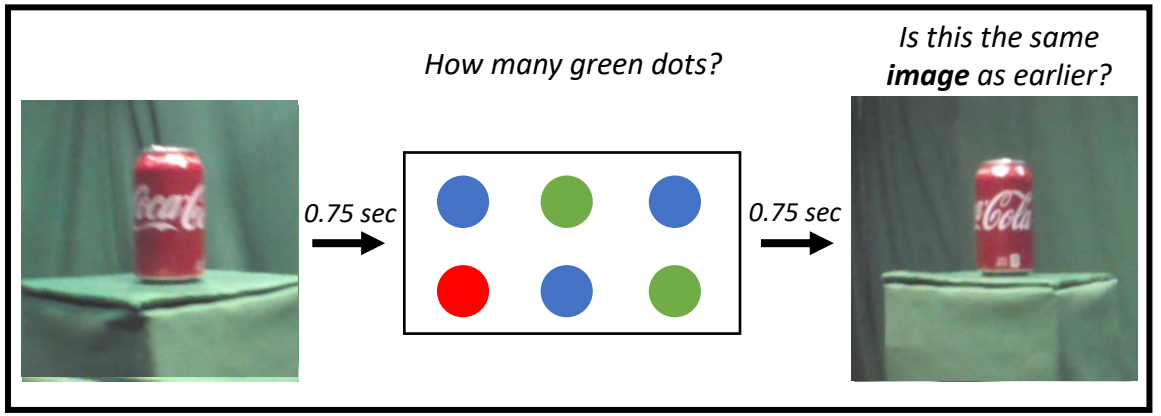
Viewpoint Variation Attack

Human Imperception Test Procedure

Imperceptible Test, Camera Shake

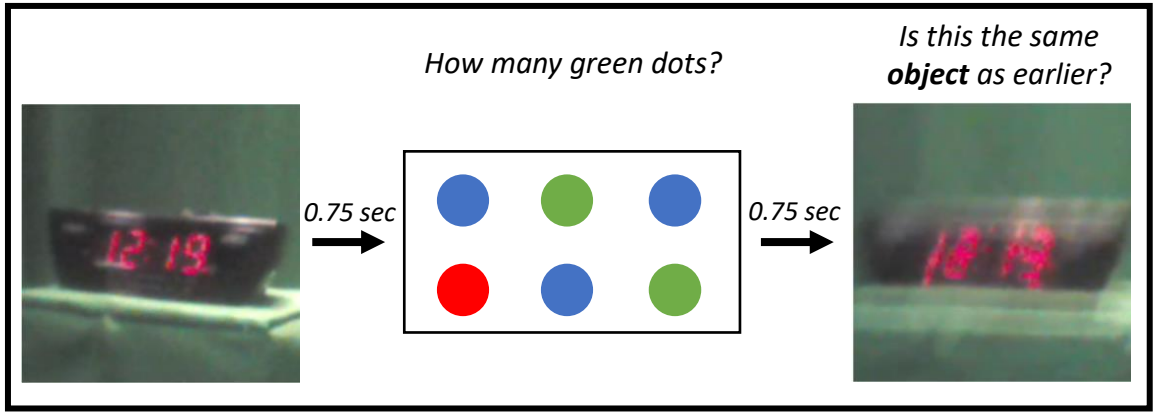


Imperceptible Test, Viewpoint Variation

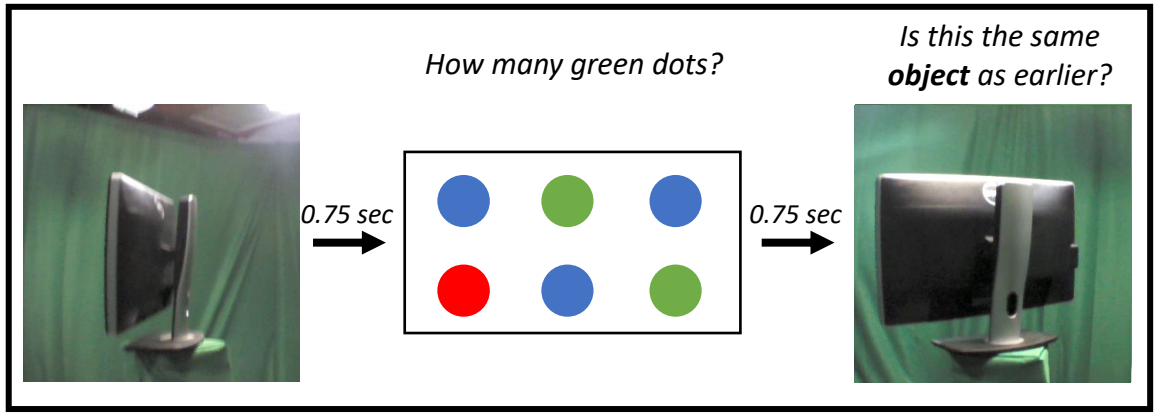


*If annotator answers “yes”, we deem image **pair** as **imperceptible**

Semantically Imperceptible Test, Camera Shake

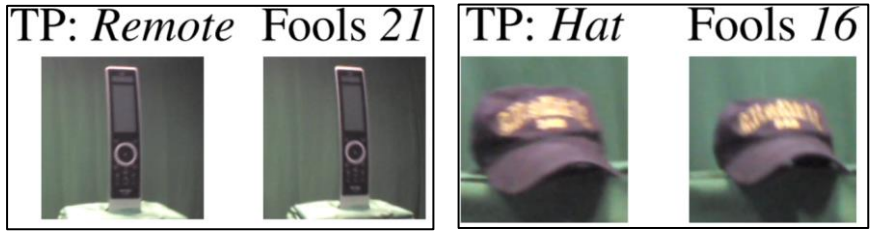


Semantically Imperceptible Test, Viewpoint Variation

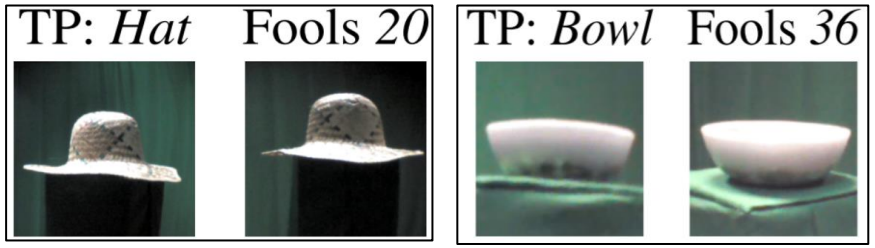


Imperceptible Perturbations

Camera Shake

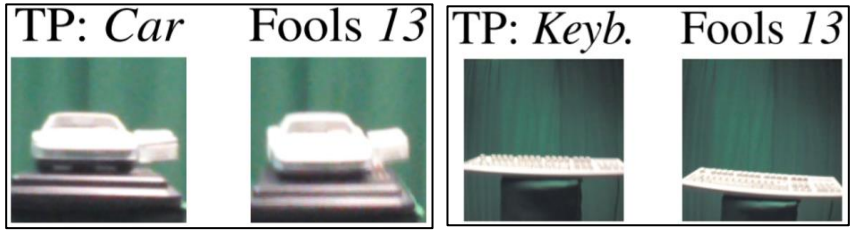


Viewpoint Variation

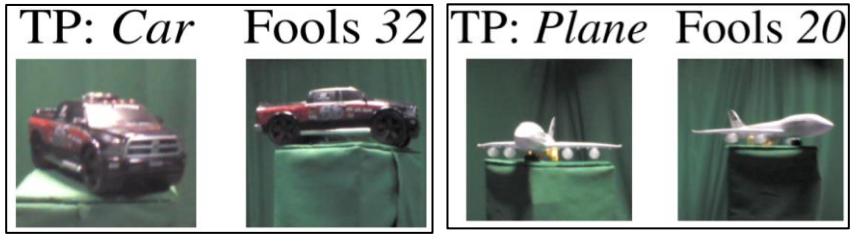


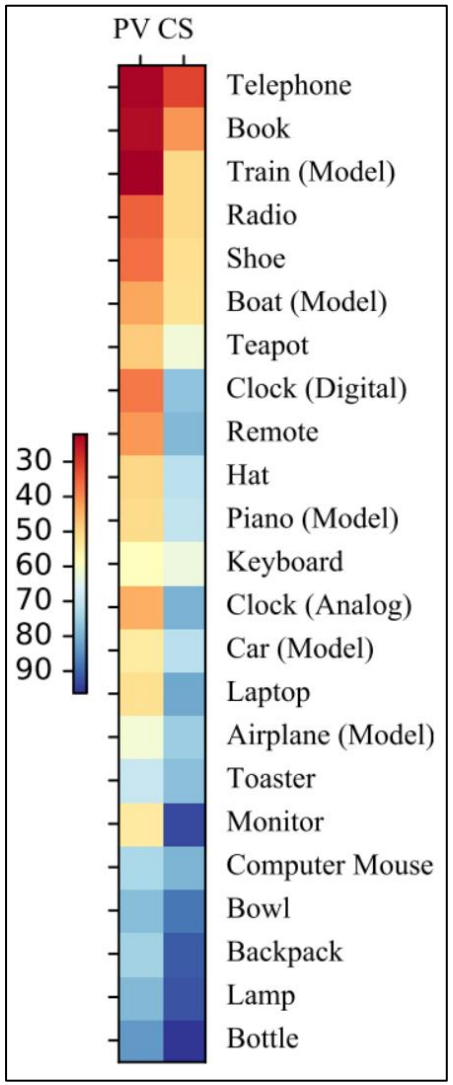
Semantically Imperceptible Perturbations

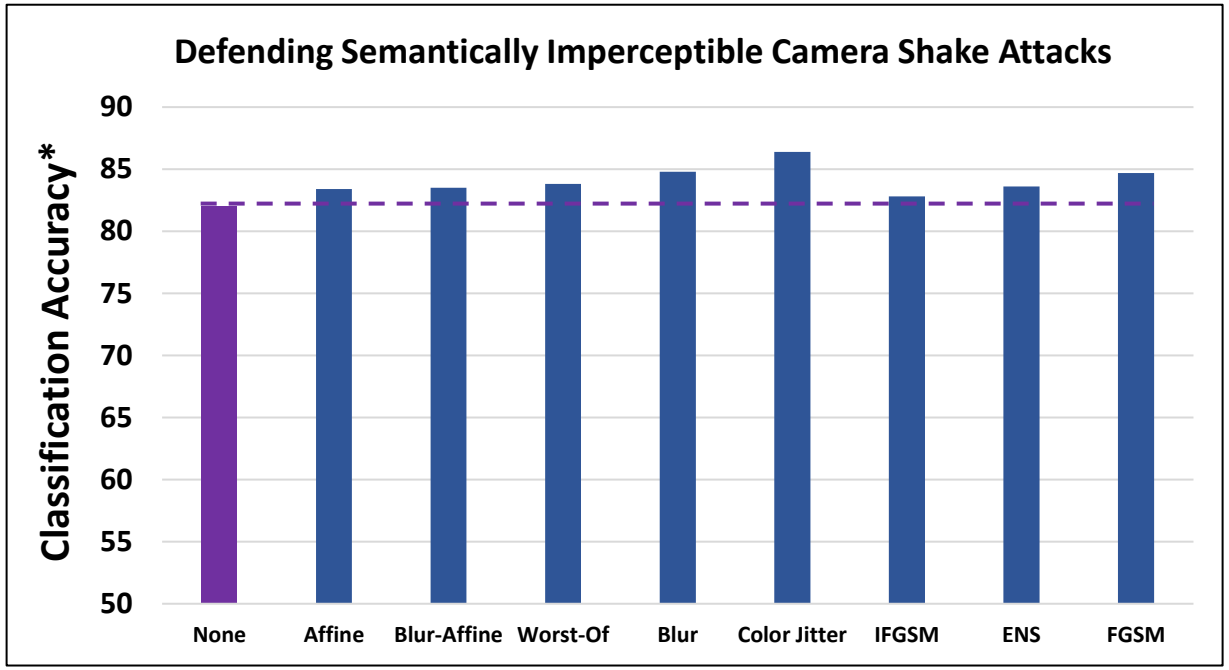
Camera Shake

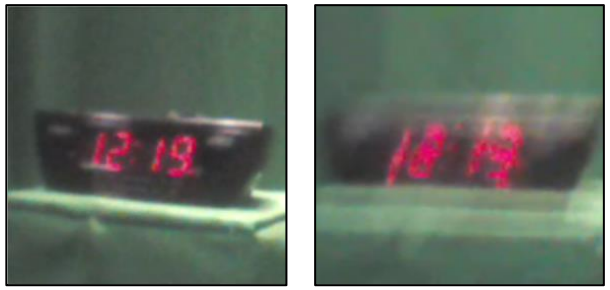


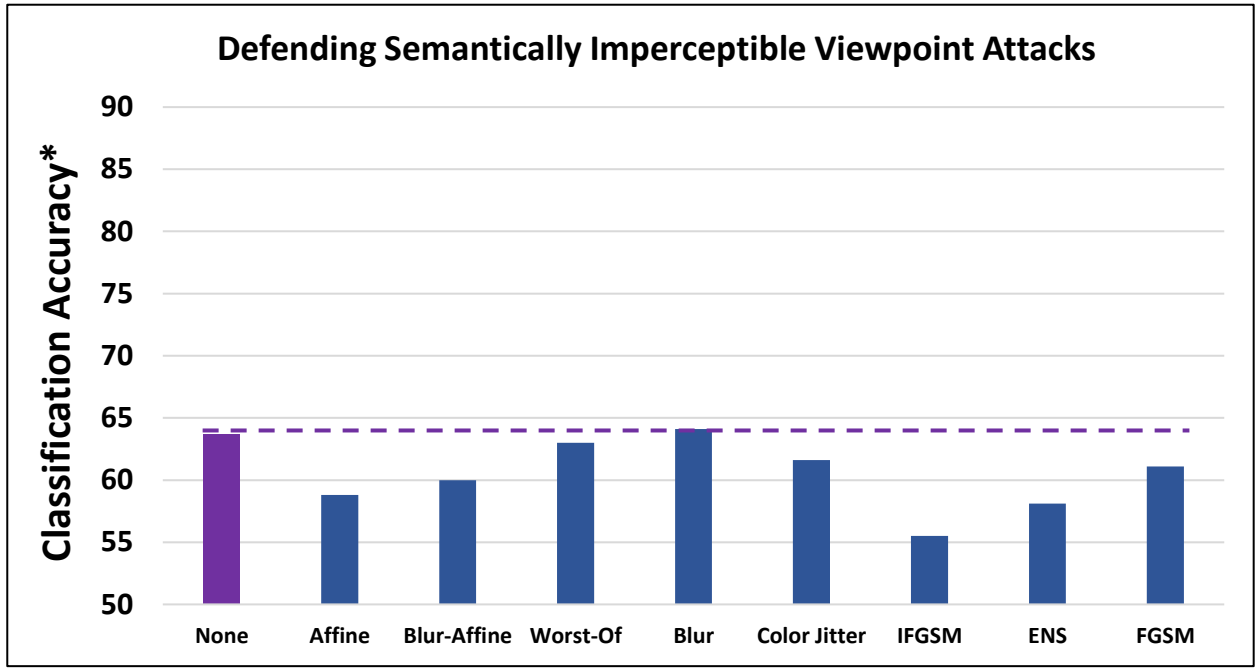
Viewpoint Variation



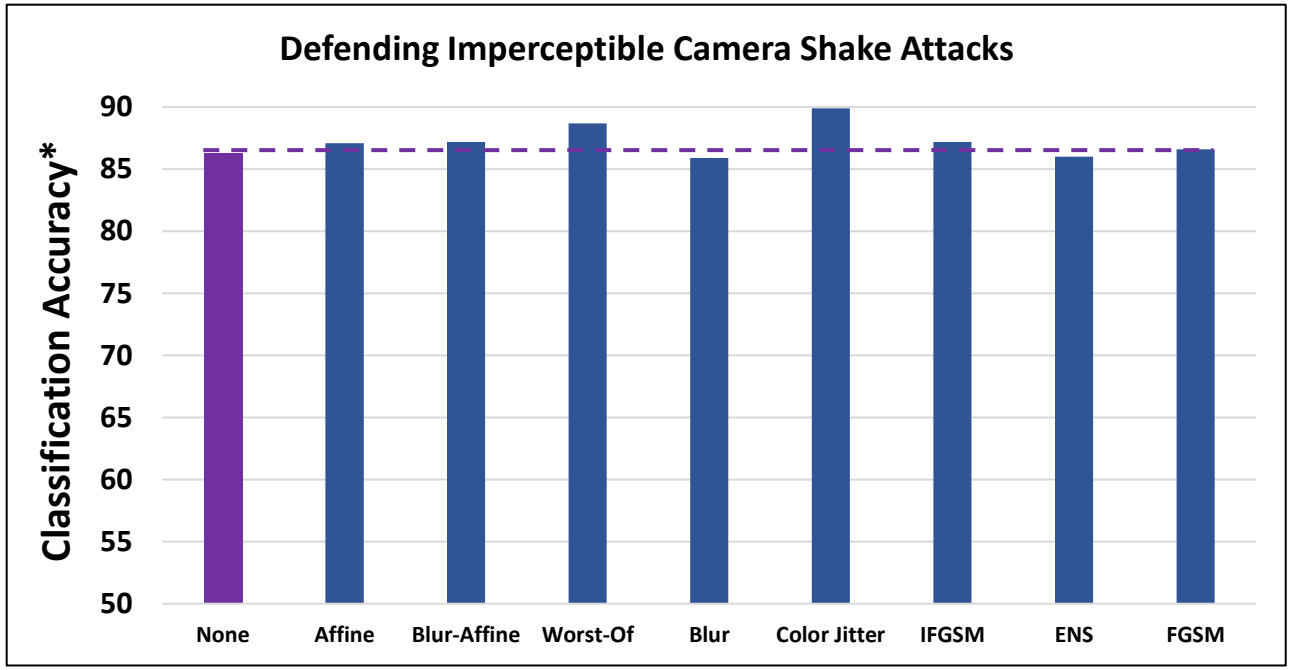


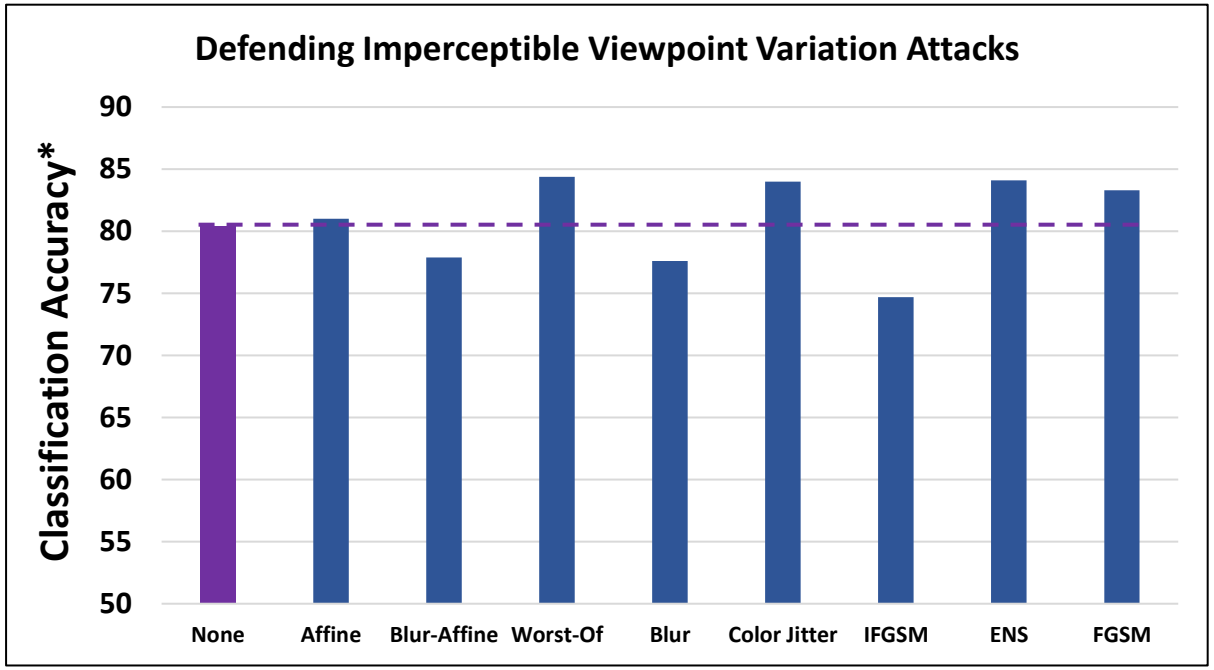


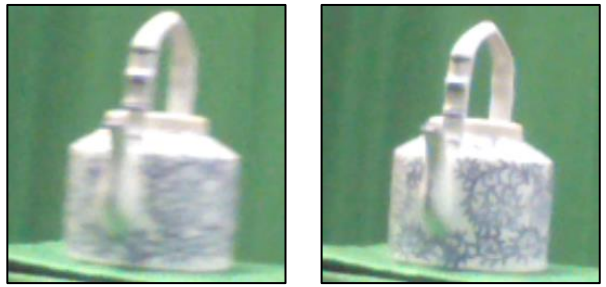




** Averaged over AlexNet, ResNet, and VGG (ImageNet pretrained, fine-tuned on OOWL frontal viewpoint training set)*







** Averaged over AlexNet, ResNet, and VGG (ImageNet pretrained, fine-tuned on OOWL frontal viewpoint training set)*